UNITED STATES PATENT APPLICATION

TITLE: PLANT TRANSCRIPTIONAL REGULATORS OF ABIOTIC STRESS

INVENTORS: **HEARD, Jacqueline E.**

KEDDIE, James S.

CREELMAN, Robert A.

PINEDA, Omaira

JIANG, Cai-Zhong

RATCLIFFE, Oliver

KUMIMOTO, Roderick W.

GUTTERSON, Neal

SHERMAN, Bradley K.

"Express Mail" Label No.:EV 059357296 US

Date of Deposit: September 30, 2003

I hereby certify under 37 C.F.R 1.10 that this correspondence is being deposited with the United States Postal Service as "Express Mail Post Office to Addressee" with sufficient postage on the date indicated above and is addressed to :

MAIL STOP Patent Application

Commissioner for Patents

PO Box 1450

Alexandria, VA 22313-1450

(Signature)

JEFFREY M. LIBBY

Printed Name)

PLANT TRANSCRIPTIONAL REGULATORS OF ABIOTIC STRESS

RELATIONSHIP TO COPENDING APPLICATIONS

This application claims the benefit of U.S. Application No. 10/412,699, filed April 10, 2003, which in turn claims the benefit of U.S. Non-provisional Application No. 09/533,030, filed March 22, 2000, which in turn claims the benefit of U.S. Provisional Application No. 60/125,814, filed March 23, 1999, U.S. Non-provisional Application 09/713,994, filed November 16, 2000, which in turn claims the benefit of U.S. Provisional Application No. 60/166,228, filed November 17, 1999, U.S. Provisional Application No. 60/197,899, filed April 17, 2000, and U.S. Provisional Application No. 60/227,439, filed August 22, 2000; U.S. Non-provisional Application No. 10/112,887, filed March 18, 2002; U.S. Non-provisional Application No. 10/286,264, filed January 23, 2003; U.S. Non-provisional Application No. 10/225,068, filed August 9, 2002; U.S. Non-provisional Application No. 10/225,066, filed August 9, 2002; U.S. Non-provisional Application No. 10/374,780, filed February 25, 2003, which claims the benefit of U.S. Non-provisional Application No. 09/837,944, filed April 18, 2001, U.S. Non-provisional Application No. 10/171,468, filed June 14, 2002, U.S. Provisional Application No. 60/310,847, filed August 9, 2001, and U.S. Provisional Application No. 60/336,049, filed November 19, 2001; U.S. Non-provisional Application "Polynucleotides and Polypeptides in Plants", filed September 18, 2003, claims the benefit of U.S. Provisional Application No. 60/434,166, filed December 17, 2002, and U.S. Provisional Application No. 60/411,837, filed September 18, 2002. The entire contents of all of these applications are hereby incorporated by reference.

FIELD OF THE INVENTION

The present invention relates to compositions and methods for modifying a plant phenotypically, said plant having altered sugar sensing and an altered response to abiotic stresses, including osmotic stresses, including germination in cold and heat, increased tolerance to drought and high salt stress.

BACKGROUND OF THE INVENTION

A plant's traits, such as its biochemical, developmental, or phenotypic characteristics, may be controlled through a number of cellular processes. One important way to manipulate that control is through transcription factors, proteins that influence the expression of a particular gene or sets of genes. Transformed and transgenic plants that comprise cells having altered levels of at least one

selected transcription factor, for example, possess advantageous or desirable traits. Strategies for manipulating traits by altering a plant cell's transcription factor content can therefore result in plants and crops with new and/or improved commercially valuable properties.

Transcription factors can modulate gene expression, either increasing or decreasing (inducing or repressing) the rate of transcription. This modulation results in differential levels of gene expression at various developmental stages, in different tissues and cell types, and in response to different exogenous (e.g., environmental) and endogenous stimuli throughout the life cycle of the organism.

Phylogenetic relationships among organisms have been demonstrated many times, and studies from a diversity of prokaryotic and eukaryotic organisms suggest a more or less gradual evolution of biochemical and physiological mechanisms and metabolic pathways. Despite different evolutionary pressures, proteins that regulate the cell cycle in yeast, plant, nematode, fly, rat, and man have common chemical or structural features and modulate the same general cellular activity. Comparisons of *Arabidopsis* gene sequences with those from other organisms where the structure and/or function may be known allow researchers to draw analogies and to develop model systems for testing hypotheses. These model systems are of great importance in developing and testing plant varieties with novel traits that may have an impact upon agronomy.

Because transcription factors are key controlling elements of biological pathways, altering the expression levels of one or more transcription factors can change entire biological pathways in an organism. For example, manipulation of the levels of selected transcription factors may result in increased expression of economically useful proteins or biomolecules in plants or improvement in other agriculturally relevant characteristics. Conversely, blocked or reduced expression of a transcription factor may reduce biosynthesis of unwanted compounds or remove an undesirable trait. Therefore, manipulating transcription factor levels in a plant offers tremendous potential in agricultural biotechnology for modifying a plant's traits, including traits that improve a plant's survival and yield during periods of abiotic stress, including germination in cold and hot conditions, and osmotic stress, including drought, salt stress, and other abiotic stresses, as noted below.

Problems associated with drought. A drought is a period of abnormally dry weather that persists long enough to produce a serious hydrologic imbalance (for example crop damage, water supply shortage, etc.). While much of the weather that we experience is brief and short-lived, drought is a more gradual phenomenon, slowly taking hold of an area and tightening its grip with time. In severe cases, drought can last for many years and can have devastating effects on agriculture and water supplies. With burgeoning population and chronic shortage of available fresh water, drought is not only the number one weather related problem in agriculture, it also ranks as one of the major natural disasters of all time, causing not only economic damage, but also loss of human lives. For example,

losses from the US drought of 1988 exceeded $40 billion, exceeding the losses caused by Hurricane Andrew in 1992, the Mississippi River floods of 1993, and the San Francisco earthquake in 1989. In some areas of the world, the effects of drought can be far more severe. In the Horn of Africa the 1984–1985 drought led to a famine that killed 750,000 people.

Problems for plants caused by low water availability include mechanical stresses caused by the withdrawal of cellular water. Drought also causes plants to become more susceptible to various diseases (Simpson (1981). "The Value of Physiological Knowledge of Water Stress in Plants", In Water Stress on Plants, (Simpson, G. M., ed.), Praeger, NY, pp. 235-265).

In addition to the many land regions of the world that are too arid for most if not all crop plants, overuse and over-utilization of available water is resulting in an increasing loss of agriculturally-usable land, a process which, in the extreme, results in desertification. The problem is further compounded by increasing salt accumulation in soils, as described above, which adds to the loss of available water in soils.

Problems associated with high salt levels. One in five hectares of irrigated land is damaged by salt, an important historical factor in the decline of ancient agrarian societies. This condition is only expected to worsen, further reducing the availability of arable land and crop production, since none of the top five food crops - wheat, corn, rice, potatoes, and soybean - can tolerate excessive salt.

Detrimental effects of salt on plants are a consequence of both water deficit resulting in osmotic stress (similar to drought stress) and the effects of excess sodium ions on critical biochemical processes. As with freezing and drought, high saline causes water deficit; the presence of high salt makes it difficult for plant roots to extract water from their environment (Buchanan et al. (2000) in Biochemistry and Molecular Biology of Plants, American Society of Plant Physiologists, Rockville, MD). Soil salinity is thus one of the more important variables that determines where a plant may thrive. In many parts of the world, sizable land areas are uncultivable due to naturally high soil salinity. To compound the problem, salination of soils that are used for agricultural production is a significant and increasing problem in regions that rely heavily on agriculture. The latter is compounded by over-utilization, over-fertilization and water shortage, typically caused by climatic change and the demands of increasing population. Salt tolerance is of particular importance early in a plant's lifecycle, since evaporation from the soil surface causes upward water movement, and salt accumulates in the upper soil layer where the seeds are placed. Thus, germination normally takes place at a salt concentration much higher than the mean salt level in the whole soil profile.

Problems associated with excessive heat. Germination of many crops is very sensitive to temperature. A transcription factor that would enhance germination in hot conditions would be useful for crops that are planted late in the season or in hot climates. Seedlings and mature plants that are

exposed to excess heat may experience heat shock, which may arise in various organs, including leaves and particularly fruit, when transpiration is insufficient to overcome heat stress. Heat also damages cellular structures, including organelles and cytoskeleton, and impairs membrane function (Buchanan et al. (2000) in Biochemistry and Molecular Biology of Plants, American Society of Plant Physiologists, Rockville, MD).

Heat shock may produce a decrease in overall protein synthesis, accompanied by expression of heat shock proteins. Heat shock proteins function as chaperones and are involved in refolding proteins denatured by heat.

Heat stress often accompanies conditions of low water availability. Heat itself is seen as an interacting stress and adds to the detrimental effects caused by water deficit conditions. Evaporative demand exhibits near exponential increases with increases in daytime temperatures and can result in high transpiration rates and low plant water potentials (Hall et al. (2000) *Plant Physiol.* 123: 1449-1458). High-temperature damage to pollen almost always occurs in conjunction with drought stress, and rarely occurs under well-watered conditions. Thus, separating the effects of heat and drought stress on pollination is difficult. Combined stress can alter plant metabolism in novel ways; therefore understanding the interaction between different stresses may be important for the development of strategies to enhance stress tolerance by genetic manipulation.

Problems associated with excessive chilling conditions. The term "chilling sensitivity" has been used to describe many types of physiological damage produced at low, but above freezing, temperatures. Most crops of tropical origins, such as soybean, rice, maize and cotton are easily damaged by chilling. Typical chilling damage includes wilting, necrosis, chlorosis or leakage of ions from cell membranes. The underlying mechanisms of chilling sensitivity are not completely understood yet, but probably involve the level of membrane saturation and other physiological deficiencies. For example, photoinhibition of photosynthesis (disruption of photosynthesis due to high light intensities) often occurs under clear atmospheric conditions subsequent to cold late summer/autumn nights. For example, chilling may lead to yield losses and lower product quality through the delayed ripening of maize. Another consequence of poor growth is the rather poor ground cover of maize fields in spring, often resulting in soil erosion, increased occurrence of weeds, and reduced uptake of nutrients. A retarded uptake of mineral nitrogen could also lead to increased losses of nitrate into the ground water. By some estimates, chilling accounts for monetary losses in the United States (US) behind only to drought and flooding.

Desirability of altered sugar sensing. Sugars are key regulatory molecules that affect diverse processes in higher plants including germination, growth, flowering, senescence, sugar metabolism and photosynthesis. Sucrose, for example, is the major transport form of photosynthate and its flux through

cells has been shown to affect gene expression and alter storage compound accumulation in seeds (source-sink relationships). Glucose-specific hexose-sensing has also been described in plants and is implicated in cell division and repression of "famine" genes (photosynthetic or glyoxylate cycles).

Water deficit is a common component of many plant stresses. Water deficit occurs in plant cells when the whole plant transpiration rate exceeds the water uptake. In addition to drought, other stresses, such as salinity and low temperature, produce cellular dehydration (McCue and Hanson (1990) *Trends Biotechnol.* 8: 358-362

Salt and drought stress signal transduction consist of ionic and osmotic homeostasis signaling pathways. The ionic aspect of salt stress is signaled via the SOS pathway where a calcium-responsive SOS3-SOS2 protein kinase complex controls the expression and activity of ion transporters such as SOS1. The pathway regulating ion homeostasis in response to salt stress has been reviewed recently by Xiong and Zhu (2002) *Plant Cell Environ.* 25: 131-139.

The osmotic component of salt stress involves complex plant reactions that overlap with drought and/or cold stress responses.

Common aspects of drought, cold and salt stress response have been reviewed recently by Xiong and Zhu (2002) *supra*). Those include:

(a) transient changes in the cytoplasmic calcium levels very early in the signaling event (Knight, (2000) *Int. Rev. Cytol.* 195: 269-324; Sanders et al. (1999) *Plant Cell* 11: 691-706);

(b) signal transduction via mitogen-activated and/or calcium dependent protein kinases (CDPKs; see Xiong et al., 2002) and protein phosphatases (Merlot et al. (2001) *Plant J.* 25: 295-303; Tähtiharju and Palva (2001) *Plant J.* 26: 461-470) ;

(c) increases in abscisic acid levels in response to stress triggering a subset of responses (Xiong et al. (2002) *supra*, and references therein) ;

(d) inositol phosphates as signal molecules (at least for a subset of the stress responsive transcriptional changes (Xiong et al. (2001) *Genes Dev.* 15: 1971-1984);

(e) activation of phospholipases which in turn generate a diverse array of second messenger molecules, some of which might regulate the activity of stress responsive kinases (phospholipase D functions in an ABA independent pathway, Frank et al. (2000) *Plant Cell* 12: 111-124);

(f) induction of late embryogenesis abundant (LEA) type genes including the CRT/DRE responsive COR/RD genes (Xiong and Zhu (2002) *supra*);

(g) increased levels of antioxidants and compatible osmolytes such as proline and soluble sugars (Hasegawa et al. (2000) *Annu. Rev. Plant Mol. Plant Physiol.* 51: 463-499); and

(h) accumulation of reactive oxygen species such as superoxide, hydrogen peroxide, and hydroxyl radicals (Hasegawa et al. (2000) *supra*).

Abscisic acid biosynthesis is regulated by osmotic stress at multiple steps. Both ABA-dependent and -independent osmotic stress signaling first modify constitutively expressed transcription factors, leading to the expression of early response transcriptional activators, which then activate downstream stress tolerance effector genes.

Based on the commonality of many aspects of cold, drought and salt stress responses, it can be concluded that genes that increase tolerance to cold or salt stress can also improve drought stress protection. In fact this has already been demonstrated for transcription factors (in the case of AtCBF/DREB1) and for other genes such as OsCDPK7 (Saijo et al. (2000) *Plant J.* 23: 319-327), or AVP1 (a vacuolar pyrophosphatase-proton-pump, Gaxiola et al. (2001) *Proc. Natl. Acad. Sci.* USA 98: 11444-11449).

The present invention relates to methods and compositions for producing transgenic plants with modified traits, particularly traits that address agricultural and food needs. These traits, including altered sugar sensing and tolerance to abiotic and osmotic stress (e.g.., tolerance to cold, high salt concentrations and drought), may provide significant value in that they allow the plant to thrive in hostile environments, where, for example, high or low temperature, low water availability or high salinity may limit or prevent growth of non-transgenic plants.

We have identified polynucleotides encoding transcription factors, including G482, G481, G485, G1364, G2345, G1781 and their equivalogs listed in the Sequence Listing, and structurally and functionally similar sequences, developed numerous transgenic plants using these polynucleotides, and have analyzed the plants for their tolerance to abiotic stresses, including those associated with heat, cold, or osmotic stresses such as drought and excessive salt. In so doing, we have identified important polynucleotide and polypeptide sequences for producing commercially valuable plants and crops as well as the methods for making them and using them. Other aspects and embodiments of the invention are described below and can be derived from the teachings of this disclosure as a whole.

SUMMARY OF THE INVENTION

The present invention pertains to transgenic plants that comprise a recombinant polynucleotide that includes a nucleotide sequence encoding a CCAAT transcription factor with the ability to regulate abiotic stress tolerance in a plant. The nucleotide sequence is capable of hybridizing to the complement

of the G482 polynucleotide sequence (SEQ ID NO:3) under stringent conditions consisting of hybridization (e.g., to filter-bound DNA, such as a hybridization procedure that includes the use of 6x SSC, 65° C, in two wash steps of 10 - 30 minutes in duration. The resultant transgenic plant has increased tolerance to abiotic stress as compared to a non-transformed plant.

The invention also encompasses transgenic plant that comprise a recombinant polynucleotide that includes a nucleotide sequence encoding a CCAAT transcription factor with the ability to regulate abiotic stress tolerance in a plant; the transcription factor comprising a CCAAT-box binding conserved domain that is at least 83% identical with the conserved CCAAT-box binding or "B" domain of the G482 polypeptide (SEQ ID NO: 4). This transgenic plant has increased tolerance to abiotic stress as compared to a non-transformed plant that does not overexpress the recombinant polynucleotide.

The present invention also relates to a method of using transgenic plants transformed with the presently disclosed transcription factor sequences, their complements or their variants to grow a progeny plant by crossing the transgenic plant with either itself or another plant, selecting seed that develops as a result of the crossing; and then growing the progeny plant from the seed. The progeny plant will generally express mRNA that encodes a transcription factor: that is, a DNA-binding protein that binds to a DNA regulatory sequence and regulates gene expression, such as that of a plant trait gene. The mRNA will generally be expressed at a level greater than a non-transformed plant; and the progeny plant is characterized by a change in a plant trait compared to the non-transformed plant.

The present invention also pertains to an expression cassette. The expression cassette comprises at least two elements, including: (1) a constitutive, inducible, or tissue-specific promoter; and (2) a recombinant polynucleotide having a polynucleotide sequence, or a complementary polynucleotide sequence thereof, selected from the group consisting of

(a) the G482 polynucleotide (SEQ ID NO: 3);

(b) a polynucleotide encoding the G482 polypeptide (SEQ ID NO: 4);

(c) a nucleotide sequence that hybridizes to the polynucleotide of (a) or (b) under the stringent conditions of 6X SSC and 65° C; and

(d) a nucleotide sequence that is at least 83% identical with the B domain found in the G482 polypeptide (SEQ ID NO: 4).

The invention is also characterized by a host cell that contains the aforementioned expression cassette.

The present invention also pertains to methods for increasing a plant's tolerance to abiotic stress. This is accomplished through the use of a vector that comprises a polynucleotide sequence that hybridizes over its full length to the complement of the G482 polynucleotide (SEQ ID NO:3) under the stringent conditions of hybridization to filter-bound DNA in 6x SSC at 65° C. The polynucleotide

sequence encodes a CCAAT transcription factor that has the property of SEQ ID NO:4 of regulating abiotic stress tolerance in a plant. The vector also includes regulatory elements that control expression of the polynucleotide sequence in a target plant. These regulatory elements flank the polynucleotide sequence. The target plant is then transformed with the vector, which transformation process generates a plant with increased tolerance to osmotic stress.

The invention is also directed to a method for producing a plant that has increased tolerance to one or more osmotic stresses. This method is performed by selecting a polynucleotide that encodes the G482 polypeptide (SEQ ID NO: 4), inserting either this polynucleotide or its complement into an expression cassette (for example, the expression cassette described above), introducing the expression cassette into a plant or plant cell in order to overexpress the G482 polypeptide, which thereby produces a plant having increased tolerance to osmotic stress. A plant that has this increased tolerance relative to a control plant not so transformed may then be identified and selected.

The invention further pertains to an isolated nucleic acid comprising a nucleotide sequence at least 99.6% identical to G482, SEQ ID NO: 3, wherein the expression of this nucleotide sequence results in increased abiotic stress tolerance in a plant.

Also encompassed by the invention are polypeptides encoded by isolated nucleic acids that are at least 99.6% identical to G482, vectors comprising isolated nucleic acid that are at least 99.6% identical to G482, and host cells and transgenic plants transformed with these isolated nucleic acids.

BRIEF DESCRIPTION OF THE SEQUENCE LISTING AND FIGURES

The file of this patent contains at least one drawing executed in color. Copies of this patent with color drawing(s) will be provided by the Patent and Trademark Office upon request and payment of the necessary fee.

The Sequence Listing provides exemplary polynucleotide and polypeptide sequences of the invention. The traits associated with the use of the sequences are included in the Examples.

CD-ROM1 is a read-only memory computer-readable compact disc and contains a copy of the Sequence Listing in ASCII text format. The Sequence Listing is named "MBI0022CIP.ST25.txt" and is 163 kilobytes in size. The copies of the Sequence Listing on the CD-ROM disc are hereby incorporated by reference in their entirety.

Figure 1 shows a conservative estimate of phylogenetic relationships among the orders of flowering plants (modified from Angiosperm Phylogeny Group (1998) *Ann. Missouri Bot. Gard.* 84: 1-49). Those plants with a single cotyledon (monocots) are a monophyletic clade nested within at least two major lineages of dicots; the eudicots are further divided into rosids and asterids. *Arabidopsis* is a rosid eudicot classified within the order Brassicales; rice is a member of the monocot order Poales.

Figure 1 was adapted from Daly et al. (2001) *Plant Physiol.* 127: 1328-1333.

Figure 2 shows a phylogenic dendogram depicting phylogenetic relationships of higher plant taxa, including clades containing tomato and *Arabidopsis*; adapted from Ku et al. (2000) *Proc. Natl. Acad. Sci.* 97: 9121-9126; and Chase et al. (1993) *Ann. Missouri Bot. Gard.* 80: 528-580.

Figure 3 is adapted from Kwong et al (2003) Plant Cell 15: 5-18, and shows crop orthologs identified through BLAST analysis of various L1L-related sequences. A phylogeny tree was then generated using ClustalX based on whole protein sequences showing the non-LEC1-like HAP3 clade of transcription factors (large box). This clade, also contains members from other species (for example, SEQ ID NOs: 18, 20, 24, 26, 48, 50, 52, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, and other sequences appearing in Table 5) are phylogenetically distinct from the LEC1-like proteins, some of which are also shown in Figure 3. The smaller box delineates the G482-like subclade, containing transcription factors that are structurally most closely related to G482, and in which several members have been shown to confer improved abiotic stress tolerance and/or altered flowering time characteristics.

Similar to Figure 3, Figure 4 shows the phylogenic relationship of sequences within the G482-subclade (within the smaller box) and the non-LEC1-like clade (larger box).

Figure 5 shows the domain structure of HAP3 proteins. HAP3 proteins contain an amino-terminal A domain, a central B domain, and a carboxy-terminal C domain. There may be relatively little sequence similarity between HAP3 proteins in the A and C domains. The A and C domains could thus provide a degree of specificity to each member of the HAP3 family. The B domain is the conserved region that specifies DNA binding and subunit association.

In Figures 6A-6F, the alignments of HAP3 polypeptides are presented, including G481, G482, G485, G1364, G2345, G1781 and related sequences from *Arabidopsis* aligned with soybean, rice and corn sequences, showing the B domains (indicated by the line that spans Figures 6B through 6D). Consensus residues within the listed sequences are indicated by boldface. The boldfaced residues in the consensus sequence that appears at the bottom of Figures 6A through 6C in their respective positions are uniquely found in the non-LEC1-like clade. The underlined serine residue appearing in the consensus sequence in its respective positions is uniquely found within the G482-like subclade. As discussed in greater detail below in Example IX, the residue positions indicated by the arrows in Figure 6B are associated with an alteration of flowering time when these polypeptides are overexpressed.

Figures 7A-7D show the effects of water deprivation and recovery from this treatment on *Arabidopsis* control and 35S::G481-overexpressing lines. After eight days of drought treatment overexpressing plants had a darker green and less withered appearance (Figure 7C) than those in the control group (Figure 7A). The differences in appearance between the control and G481-overexpressing plants after they were rewatered was even more striking. Most (11 of 12 plants; Figure

7B) of this set of control plants died after rewatering, indicating the inability to recover following severe water deprivation, whereas all nine of the overexpressor plants of the line shown recovered from this drought treatment (Figure 7D). The results shown in Figures 7A-7D were typical of a number of control and 35S::G481-overexpressing lines.

Figures 8A and 8B show the effects of salt stress on *Arabidopsis* seed germination. The three lines of G481-and G482 overexpressors on these two plate had longer roots and showed greater cotyledon expansion (arrows) after three days on 150 mM NaCl than the control seedlings on the right-hand sides of the plates.

In Figure 9A, G481 null mutant seedlings (labeled K481) show reduced tolerance of osmotic stress, relative to the control seedlings in Figure 8B, as evidenced by the reduced cotyledon expansion and root growth in the former group. Without salt stress tolerance on control media, (Figures 9C, G481 null mutants; and 9D, control seedlings), the knocked out and control plants appear the same.

Figures 10A-10D show the effects of stress-related treatments on G485 overexpressing seedlings (35S::G485 lines) in plate assays. In each treatment, including cold, high sucrose, high salt and ABA germination assays, the overexpressors fared much better than the wild-type controls exposed to the same treatments in Figures 10E-10H, respectively, as evidenced by the enhanced cotyledon expansion and root growth seen with the overexpressing seedlings.

Figures 11A - 11C depict the effects of G485 knockout and overexpression on flowering time and maturation. As seen in Figure 10A, a T-DNA insertion knockout mutation containing a SALK_062245 insertion was shown to flower several days later than wild-type control plants. The plants in Figure 11A are shown 44 days after germination. Figure 11C shows that G485 primary transformants flowered distinctly earlier than wild-type controls. These plants are shown 24 days after germination. These effects were observed in each of two independent T1 plantings derived from separate transformation dates. Additionally, accelerated flowering was also seen in plants that overexpressed G485 from a two component system (35S::LexA;op-LexA::G485). These studies indicated that G485 is both sufficient to act as a floral activator, and is also necessary in that role within the plant. G485 overexpressor plants also matured and set siliques much more rapidly than wild type controls, as shown in Figure 11B with plants 39 days post-germination.

DESCRIPTION OF THE INVENTION

In an important aspect, the present invention relates to polynucleotides and polypeptides, for example, for modifying phenotypes of plants, particularly those associated with osmotic stress tolerance. Throughout this disclosure, various information sources are referred to and/or are

specifically incorporated. The information sources include scientific journal articles, patent documents, textbooks, and World Wide Web browser-inactive page addresses, for example. While the reference to these information sources clearly indicates that they can be used by one of skill in the art, each and every one of the information sources cited herein are specifically incorporated in their entirety, whether or not a specific mention of "incorporation by reference" is noted. The contents and teachings of each and every one of the information sources can be relied on and used to make and use embodiments of the invention.

As used herein and in the appended claims, the singular forms "a," "an," and "the" include plural reference unless the context clearly dictates otherwise. Thus, for example, a reference to "a plant" includes a plurality of such plants, and a reference to "a stress" is a reference to one or more stresses and equivalents thereof known to those skilled in the art, and so forth.

DEFINITIONS

"Nucleic acid molecule" refers to a oligonucleotide, polynucleotide or any fragment thereof. It may be DNA or RNA of genomic or synthetic origin, double-stranded or single-stranded, and combined with carbohydrate, lipids, protein, or other materials to perform a particular activity such as transformation or form a useful composition such as a peptide nucleic acid (PNA).

"Polynucleotide" is a nucleic acid molecule comprising a plurality of polymerized nucleotides, e.g., at least about 15 consecutive polymerized nucleotides, optionally at least about 30 consecutive nucleotides, at least about 50 consecutive nucleotides. A polynucleotide may be a nucleic acid, oligonucleotide, nucleotide, or any fragment thereof. In many instances, a polynucleotide comprises a nucleotide sequence encoding a polypeptide (or protein) or a domain or fragment thereof. Additionally, the polynucleotide may comprise a promoter, an intron, an enhancer region, a polyadenylation site, a translation initiation site, 5' or 3' untranslated regions, a reporter gene, a selectable marker, or the like. The polynucleotide can be single stranded or double stranded DNA or RNA. The polynucleotide optionally comprises modified bases or a modified backbone. The polynucleotide can be, e.g., genomic DNA or RNA, a transcript (such as an mRNA), a cDNA, a PCR product, a cloned DNA, a synthetic DNA or RNA, or the like. The polynucleotide can be combined with carbohydrate, lipids, protein, or other materials to perform a particular activity such as transformation or form a useful composition such as a peptide nucleic acid (PNA). The polynucleotide can comprise a sequence in either sense or antisense orientations. "Oligonucleotide" is substantially equivalent to the terms amplimer, primer, oligomer, element, target, and probe and is preferably single stranded.

"Gene" or "gene sequence" refers to the partial or complete coding sequence of a gene, its

complement, and its 5' or 3' untranslated regions. A gene is also a functional unit of inheritance, and in physical terms is a particular segment or sequence of nucleotides along a molecule of DNA (or RNA, in the case of RNA viruses) involved in producing a polypeptide chain. The latter may be subjected to subsequent processing such as splicing and folding to obtain a functional protein or polypeptide.. A gene may be isolated, partially isolated, or be found with an organism's genome. By way of example, a transcription factor gene encodes a transcription factor polypeptide, which may be functional or require processing to function as an initiator of transcription.

Operationally, genes may be defined by the cis-trans test, a genetic test that determines whether two mutations occur in the same gene and which may be used to determine the limits of the genetically active unit (Rieger et al. (1976) Glossary of Genetics and Cytogenetics: Classical and Molecular, 4th ed., Springer Verlag. Berlin). A gene generally includes regions preceding ("leaders"; upstream) and following ("trailers"; downstream) of the coding region. A gene may also include intervening, non-coding sequences, referred to as "introns", located between individual coding segments, referred to as "exons". Most genes have an associated promoter region, a regulatory sequence 5' of the transcription initiation codon (there are some genes that do not have an identifiable promoter). The function of a gene may also be regulated by enhancers, operators, and other regulatory elements.

A "recombinant polynucleotide" is a polynucleotide that is not in its native state, e.g., the polynucleotide comprises a nucleotide sequence not found in nature, or the polynucleotide is in a context other than that in which it is naturally found, e.g., separated from nucleotide sequences with which it typically is in proximity in nature, or adjacent (or contiguous with) nucleotide sequences with which it typically is not in proximity. For example, the sequence at issue can be cloned into a vector, or otherwise recombined with one or more additional nucleic acid.

An "isolated polynucleotide" is a polynucleotide whether naturally occurring or recombinant, that is present outside the cell in which it is typically found in nature, whether purified or not. Optionally, an isolated polynucleotide is subject to one or more enrichment or purification procedures, e.g., cell lysis, extraction, centrifugation, precipitation, or the like.

A "polypeptide" is an amino acid sequence comprising a plurality of consecutive polymerized amino acid residues e.g., at least about 15 consecutive polymerized amino acid residues, optionally at least about 30 consecutive polymerized amino acid residues, at least about 50 consecutive polymerized amino acid residues. In many instances, a polypeptide comprises a polymerized amino acid residue sequence that is a transcription factor or a domain or portion or fragment thereof. Additionally, the polypeptide may comprise 1) a localization domain, 2) an activation domain, 3) a repression domain, 4) an oligomerization domain, or 5) a DNA-binding domain, or the like. The polypeptide optionally

comprises modified amino acid residues, naturally occurring amino acid residues not encoded by a codon, non-naturally occurring amino acid residues.

"Protein" refers to an amino acid sequence, oligopeptide, peptide, polypeptide or portions thereof whether naturally occurring or synthetic.

"Portion", as used herein, refers to any part of a protein used for any purpose, but especially for the screening of a library of molecules which specifically bind to that portion or for the production of antibodies.

A "recombinant polypeptide" is a polypeptide produced by translation of a recombinant polynucleotide. A "synthetic polypeptide" is a polypeptide created by consecutive polymerization of isolated amino acid residues using methods well known in the art. An "isolated polypeptide," whether a naturally occurring or a recombinant polypeptide, is more enriched in (or out of) a cell than the polypeptide in its natural state in a wild-type cell, e.g., more than about 5% enriched, more than about 10% enriched, or more than about 20%, or more than about 50%, or more, enriched, i.e., alternatively denoted: 105%, 110%, 120%, 150% or more, enriched relative to wild type standardized at 100%. Such an enrichment is not the result of a natural response of a wild-type plant. Alternatively, or additionally, the isolated polypeptide is separated from other cellular components with which it is typically associated, e.g., by any of the various protein purification methods herein.

"Homology" refers to sequence similarity between a reference sequence and at least a fragment of a newly sequenced clone insert or its encoded amino acid sequence.

"Hybridization complex" refers to a complex between two nucleic acid molecules by virtue of the formation of hydrogen bonds between purines and pyrimidines.

"Identity" or "similarity" refers to sequence similarity between two polynucleotide sequences or between two polypeptide sequences, with identity being a more strict comparison. The phrases "percent identity" and "% identity" refer to the percentage of sequence similarity found in a comparison of two or more polynucleotide sequences or two or more polypeptide sequences. "Sequence similarity" refers to the percent similarity in base pair sequence (as determined by any suitable method) between two or more polynucleotide sequences. Two or more sequences can be anywhere from 0-100% similar, or any integer value therebetween. Identity or similarity can be determined by comparing a position in each sequence that may be aligned for purposes of comparison. When a position in the compared sequence is occupied by the same nucleotide base or amino acid, then the molecules are identical at that position. A degree of similarity or identity between polynucleotide sequences is a function of the number of identical or matching nucleotides at positions shared by the polynucleotide sequences. A degree of identity of polypeptide sequences is a function of the number of identical amino acids at positions shared by the polypeptide sequences. A degree of homology or similarity of polypeptide

sequences is a function of the number of amino acids at positions shared by the polypeptide sequences.

The term "amino acid consensus motif" refers to the portion or subsequence of a polypeptide sequence that is substantially conserved among the polypeptide transcription factors listed in the Sequence Listing.

"Alignment" refers to a number of nucleotide bases or amino acid residue sequences aligned by lengthwise comparison so that components in common (i.e., nucleotide bases or amino acid residues) may be visually and readily identified. The fraction or percentage of components in common is related to the homology or identity between the sequences. Alignments such as those of Figures 6A-6F may be used to identify conserved domains and relatedness within these domains. An alignment may suitably be determined by means of computer programs known in the art, such as MACVECTOR software (1999) (Accelrys, Inc., San Diego, CA).

A "conserved domain" or "conserved region" as used herein refers to a region in heterologous polynucleotide or polypeptide sequences where there is a relatively high degree of sequence identity between the distinct sequences. A CCAAT-box binding conserved domain, such as one of the domains shown in Table 1, is an example of a conserved domain.

With respect to polynucleotides encoding presently disclosed transcription factors, a conserved domain is preferably at least 10 base pairs (bp) in length.

A "conserved domain", with respect to presently disclosed polypeptides refers to a domain within a transcription factor family that exhibits a higher degree of sequence homology, such as at least 26% sequence similarity, at least 16% sequence identity, preferably at least 40% sequence identity, preferably at least 65% sequence identity including conservative substitutions, and more preferably at least 80% sequence identity, and even more preferably at least 85%, or at least about 86%, or at least about 87%, or at least about 88%, or at least about 90%, or at least about 95%, or at least about 98% amino acid residue sequence identity of a polypeptide of consecutive amino acid residues. A fragment or domain can be referred to as outside a conserved domain, outside a consensus sequence, or outside a consensus DNA-binding site that is known to exist or that exists for a particular transcription factor class, family, or sub-family. In this case, the fragment or domain will not include the exact amino acids of a consensus sequence or consensus DNA-binding site of a transcription factor class, family or sub-family, or the exact amino acids of a particular transcription factor consensus sequence or consensus DNA-binding site. Furthermore, a particular fragment, region, or domain of a polypeptide, or a polynucleotide encoding a polypeptide, can be "outside a conserved domain" if all the amino acids of the fragment, region, or domain fall outside of a defined conserved domain(s) for a polypeptide or protein. Sequences having lesser degrees of identity but comparable biological activity are considered to be equivalents.

As one of ordinary skill in the art recognizes, conserved domains may be identified as regions or domains of identity to a specific consensus sequence (see, for example, Riechmann et al. (2000) *supra*). Thus, by using alignment methods well known in the art, the conserved domains of the plant transcription factors for the CAAT-element binding proteins (Forsburg and Guarente (1989) *Genes Dev.* 3: 1166-1178) may be determined.

The CCAAT-box binding conserved domains or conserved domains for SEQ ID NO: 2, 4, 6, 8 and 10 and similar sequences are listed in Table 1. Also, the polypeptides of Table 1 have CCAAT-box binding conserved domains specifically indicated by start and stop sites. A comparison of the regions of the polypeptides in Table 1 allows one of skill in the art to identify "B" or CCAAT-box binding conserved domains, or conserved domains for any of the polypeptides listed or referred to in this disclosure.

"Complementary" refers to the natural hydrogen bonding by base pairing between purines and pyrimidines. For example, the sequence A-C-G-T (5' -> 3') forms hydrogen bonds with its complements A-C-G-T (5' -> 3') or A-C-G-U (5' -> 3'). Two single-stranded molecules may be considered partially complementary, if only some of the nucleotides bond, or "completely complementary" if all of the nucleotides bond. The degree of complementarity between nucleic acid strands affects the efficiency and strength of the hybridization and amplification reactions. "Fully complementary" refers to the case where bonding occurs between every base pair and its complement in a pair of sequences, and the two sequences have the same number of nucleotides.

The terms "highly stringent" or "highly stringent condition" refer to conditions that permit hybridization of DNA strands whose sequences are highly complementary, wherein these same conditions exclude hybridization of significantly mismatched DNAs. Polynucleotide sequences capable of hybridizing under stringent conditions with the polynucleotides of the present invention may be, for example, variants of the disclosed polynucleotide sequences, including allelic or splice variants, or sequences that encode orthologs or paralogs of presently disclosed polypeptides. Nucleic acid hybridization methods are disclosed in detail by Kashima et al. (1985) *Nature* 313:402-404, and Sambrook et al. (1989) Molecular Cloning: A Laboratory Manual, 2nd Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y ("Sambrook"); and by Haymes et al., "Nucleic Acid Hybridization: A Practical Approach", IRL Press, Washington, D.C. (1985), which references are incorporated herein by reference.

In general, stringency is determined by the temperature, ionic strength, and concentration of denaturing agents (e.g., formamide) used in a hybridization and washing procedure (for a more detailed description of establishing and determining stringency, see below). The degree to which two nucleic acids hybridize under various conditions of stringency is correlated with the extent of their similarity.

Thus, similar nucleic acid sequences from a variety of sources, such as within a plant's genome (as in the case of paralogs) or from another plant (as in the case of orthologs) that may perform similar functions can be isolated on the basis of their ability to hybridize with known transcription factor sequences. Numerous variations are possible in the conditions and means by which nucleic acid hybridization can be performed to isolate transcription factor sequences having similarity to transcription factor sequences known in the art and are not limited to those explicitly disclosed herein. Such an approach may be used to isolate polynucleotide sequences having various degrees of similarity with disclosed transcription factor sequences, such as, for example, encoded transcription factors having 60% or greater identity with disclosed transcription factors, or 83% or greater identity with the B domain of disclosed transcription factors.

Regarding the terms "paralog" and "ortholog", homologous polynucleotide sequences and homologous polypeptide sequences may be paralogs or orthologs of the claimed polynucleotide or polypeptide sequence. Orthologs and paralogs are evolutionarily related genes that have similar sequence and similar functions. Orthologs are structurally related genes in different species that are derived by a speciation event. Paralogs are structurally related genes within a single species that are derived by a duplication event. Sequences that are sufficiently similar to one another will be appreciated by those of skill in the art and may be based upon percentage identity of the complete sequences, percentage identity of a conserved domain or sequence within the complete sequence, percentage similarity to the complete sequence, percentage similarity to a conserved domain or sequence within the complete sequence, and/or an arrangement of contiguous nucleotides or peptides particular to a conserved domain or complete sequence. Sequences that are sufficiently similar to one another will also bind in a similar manner to the same DNA binding sites of transcriptional regulatory elements using methods well known to those of skill in the art.

The term "equivalog" describes members of a set of homologous proteins that are conserved with respect to function since their last common ancestor. Related proteins are grouped into equivalog families, and otherwise into protein families with other hierarchically defined homology types. This definition is provided at the Institute for Genomic Research (TIGR) world wide web (www) website, " tigr.org " under the heading "Terms associated with TIGRFAMs".

The term "variant", as used herein, may refer to polynucleotides or polypeptides, that differ from the presently disclosed polynucleotides or polypeptides, respectively, in sequence from each other, and as set forth below.

With regard to polynucleotide variants, differences between presently disclosed polynucleotides and polynucleotide variants are limited so that the nucleotide sequences of the former and the latter are closely similar overall and, in many regions, identical. Due to the degeneracy of the

genetic code, differences between the former and latter nucleotide sequences o may be silent (i.e., the amino acids encoded by the polynucleotide are the same, and the variant polynucleotide sequence encodes the same amino acid sequence as the presently disclosed polynucleotide. Variant nucleotide sequences may encode different amino acid sequences, in which case such nucleotide differences will result in amino acid substitutions, additions, deletions, insertions, truncations or fusions with respect to the similar disclosed polynucleotide sequences. These variations result in polynucleotide variants encoding polypeptides that share at least one functional characteristic. The degeneracy of the genetic code also dictates that many different variant polynucleotides can encode identical and/or substantially similar polypeptides in addition to those sequences illustrated in the Sequence Listing.

Also within the scope of the invention is a variant of a transcription factor nucleic acid listed in the Sequence Listing, that is, one having a sequence that differs from the one of the polynucleotide sequences in the Sequence Listing, or a complementary sequence, that encodes a functionally equivalent polypeptide (i.e., a polypeptide having some degree of equivalent or similar biological activity) but differs in sequence from the sequence in the Sequence Listing, due to degeneracy in the genetic code. Included within this definition are polymorphisms that may or may not be readily detectable using a particular oligonucleotide probe of the polynucleotide encoding polypeptide, and improper or unexpected hybridization to allelic variants, with a locus other than the normal chromosomal locus for the polynucleotide sequence encoding polypeptide.

"Allelic variant" or "polynucleotide allelic variant" refers to any of two or more alternative forms of a gene occupying the same chromosomal locus. Allelic variation arises naturally through mutation, and may result in phenotypic polymorphism within populations. Gene mutations may be "silent" or may encode polypeptides having altered amino acid sequence. "Allelic variant" and "polypeptide allelic variant" may also be used with respect to polypeptides, and in this case the term refer to a polypeptide encoded by an allelic variant of a gene.

"Splice variant" or "polynucleotide splice variant" as used herein refers to alternative forms of RNA transcribed from a gene. Splice variation naturally occurs as a result of alternative sites being spliced within a single transcribed RNA molecule or between separately transcribed RNA molecules, and may result in several different forms of mRNA transcribed from the same gene. This, splice variants may encode polypeptides having different amino acid sequences, which may or may not have similar functions in the organism. "Splice variant" or "polypeptide splice variant" may also refer to a polypeptide encoded by a splice variant of a transcribed mRNA.

As used herein, "polynucleotide variants" may also refer to polynucleotide sequences that encode paralogs and orthologs of the presently disclosed polypeptide sequences. "Polypeptide variants" may refer to polypeptide sequences that are paralogs and orthologs of the presently disclosed

polypeptide sequences.

Differences between presently disclosed polypeptides and polypeptide variants are limited so that the sequences of the former and the latter are closely similar overall and, in many regions, identical. Presently disclosed polypeptide sequences and similar polypeptide variants may differ in amino acid sequence by one or more substitutions, additions, deletions, fusions and truncations, which may be present in any combination. These differences may produce silent changes and result in a functionally equivalent transcription factor. Thus, it will be readily appreciated by those of skill in the art, that any of a variety of polynucleotide sequences is capable of encoding the transcription factors and transcription factor homolog polypeptides of the invention. A polypeptide sequence variant may have "conservative" changes, wherein a substituted amino acid has similar structural or chemical properties. Deliberate amino acid substitutions may thus be made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues, as long as the functional or biological activity of the transcription factor is retained. For example, negatively charged amino acids may include aspartic acid and glutamic acid, positively charged amino acids may include lysine and arginine, and amino acids with uncharged polar head groups having similar hydrophilicity values may include leucine, isoleucine, and valine; glycine and alanine; asparagine and glutamine; serine and threonine; and phenylalanine and tyrosine (for more detail on conservative substitutions, see Table 3). More rarely, a variant may have "non-conservative" changes, e.g., replacement of a glycine with a tryptophan. Similar minor variations may also include amino acid deletions or insertions, or both. Related polypeptides may comprise, for example, additions and/or deletions of one or more N-linked or O-linked glycosylation sites, or an addition and/or a deletion of one or more cysteine residues. Guidance in determining which and how many amino acid residues may be substituted, inserted or deleted without abolishing functional or biological activity may be found using computer programs well known in the art, for example, DNASTAR software (see USPN 5,840,544).

"Ligand" refers to any molecule, agent, or compound that will bind specifically to a complementary site on a nucleic acid molecule or protein. Such ligands stabilize or modulate the activity of nucleic acid molecules or proteins of the invention and may be composed of at least one of the following: inorganic and organic substances including nucleic acids, proteins, carbohydrates, fats, and lipids.

"Modulates" refers to a change in activity (biological, chemical, or immunological) or lifespan resulting from specific binding between a molecule and either a nucleic acid molecule or a protein.

The term "plant" includes whole plants, shoot vegetative organs/structures (*e.g.*, leaves, stems and tubers), roots, flowers and floral organs/structures (for example, bracts, sepals, petals, stamens,

carpels, anthers and ovules), seed (including embryo, endosperm, and seed coat) and fruit (the mature ovary), plant tissue (for example, vascular tissue, ground tissue, and the like) and cells (for example, guard cells, egg cells, and the like), and progeny of same. The class of plants that can be used in the method of the invention is generally as broad as the class of higher and lower plants amenable to transformation techniques, including angiosperms (monocotyledonous and dicotyledonous plants), gymnosperms, ferns, horsetails, psilophytes, lycophytes, bryophytes, and multicellular algae. (See for example, Figure 1, adapted from Daly et al. (2001) *Plant Physiol.* 127: 1328-1333; Figure 2, adapted from Ku et al. (2000) *Proc. Natl. Acad. Sci.* 97: 9121-9126; and see also Tudge in The Variety of Life, Oxford University Press, New York, NY (2000) pp. 547-606).

A "transgenic plant" refers to a plant that contains genetic material not found in a wild-type plant of the same species, variety or cultivar. The genetic material may include a transgene, an insertional mutagenesis event (such as by transposon or T-DNA insertional mutagenesis), an activation tagging sequence, a mutated sequence, a homologous recombination event or a sequence modified by chimeraplasty. Typically, the foreign genetic material has been introduced into the plant by human manipulation, but any method can be used as one of skill in the art recognizes.

A transgenic plant may contain an expression vector or cassette. The expression cassette typically comprises a polypeptide-encoding sequence operably linked (i.e., under regulatory control of) to appropriate inducible or constitutive regulatory sequences that allow for the expression of polypeptide. The expression cassette can be introduced into a plant by transformation or by breeding after transformation of a parent plant. A plant refers to a whole plant as well as to a plant part, such as seed, fruit, leaf, or root, plant tissue, plant cells or any other plant material, e.g., a plant explant, as well as to progeny thereof, and to *in vitro* systems that mimic biochemical or cellular components or processes in a cell.

"Control plant" refers to a plant that serves as a standard of comparison for testing the results of a treatment or genetic alteration, or the degree of altered expression of a gene or gene product. Examples of control plants include plants that are untreated, or genetically unaltered (i.e., wild-type).

"Wild type", as used herein, refers to a cell, tissue or plant that has not been genetically modified to knock out or overexpress one or more of the presently disclosed transcription factors. Wild-type cells, tissue or plants may be used as controls to compare levels of expression and the extent and nature of trait modification with cells, tissue or plants in which transcription factor expression is altered or ectopically expressed, e.g., in that it has been knocked out or overexpressed.

"Fragment", with respect to a polynucleotide, refers to a clone or any part of a polynucleotide molecule that retains a usable, functional characteristic. Useful fragments include oligonucleotides and polynucleotides that may be used in hybridization or amplification technologies or in the regulation of

replication, transcription or translation. A polynucleotide fragment" refers to any subsequence of a polynucleotide, typically, of at least about 9 consecutive nucleotides, preferably at least about 30 nucleotides, more preferably at least about 50 nucleotides, of any of the sequences provided herein. Exemplary polynucleotide fragments are the first sixty consecutive nucleotides of the transcription factor polynucleotides listed in the Sequence Listing. Exemplary fragments also include fragments that comprise a region that encodes a B domain of a transcription factor, for example, amino acid residues 26-116 of G482 (SEQ ID NO: 4), as noted in Table 1.

Fragments may also include subsequences of polypeptides and protein molecules, or a subsequence of the polypeptide. Fragments may have uses in that they may have antigenic potential. In some cases, the fragment or domain is a subsequence of the polypeptide which performs at least one biological function of the intact polypeptide in substantially the same manner, or to a similar extent, as does the intact polypeptide. For example, a polypeptide fragment can comprise a recognizable structural motif or functional domain such as a DNA-binding site or domain that binds to a DNA promoter region, an activation domain, or a domain for protein-protein interactions, and may initiate transcription. Fragments can vary in size from as few as 3 amino acid residues to the full length of the intact polypeptide, but are preferably at least about 30 amino acid residues in length and more preferably at least about 60 amino acid residues in length. Exemplary polypeptide fragments are the first twenty consecutive amino acids of a mammalian protein encoded by are the first twenty consecutive amino acids of the transcription factor polypeptides listed in the Sequence Listing. Exemplary fragments also include fragments that comprise a B domain of a transcription factor, for example, amino acid residues 26-116 of G482 (SEQ ID NO: 4), as noted in Table 1.

The invention also encompasses production of DNA sequences that encode transcription factors and transcription factor derivatives, or fragments thereof, entirely by synthetic chemistry. After production, the synthetic sequence may be inserted into any of the many available expression vectors and cell systems using reagents well known in the art. Moreover, synthetic chemistry may be used to introduce mutations into a sequence encoding transcription factors or any fragment thereof.

"Derivative" refers to the chemical modification of a nucleic acid molecule or amino acid sequence. Chemical modifications can include replacement of hydrogen by an alkyl, acyl, or amino group or glycosylation, pegylation, or any similar process that retains or enhances biological activity or lifespan of the molecule or sequence.

A "trait" refers to a physiological, morphological, biochemical, or physical characteristic of a plant or particular plant material or cell. In some instances, this characteristic is visible to the human eye, such as seed or plant size, or can be measured by biochemical techniques, such as detecting the protein, starch, or oil content of seed or leaves, or by observation of a metabolic or physiological

process, e.g. by measuring tolerance to water deprivation or particular salt or sugar concentrations, or by the observation of the expression level of a gene or genes, e.g., by employing Northern analysis, RT-PCR, microarray gene expression assays, or reporter gene expression systems, or by agricultural observations such as osmotic stress tolerance or yield. Any technique can be used to measure the amount of, comparative level of, or difference in any selected chemical compound or macromolecule in the transgenic plants, however.

"Trait modification" refers to a detectable difference in a characteristic in a plant ectopically expressing a polynucleotide or polypeptide of the present invention relative to a plant not doing so, such as a wild-type plant. In some cases, the trait modification can be evaluated quantitatively. For example, the trait modification can entail at least about a 2% increase or decrease in an observed trait (difference), at least a 5% difference, at least about a 10% difference, at least about a 20% difference, at least about a 30%, at least about a 50%, at least about a 70%, or at least about a 100%, or an even greater difference compared with a wild-type plant. It is known that there can be a natural variation in the modified trait. Therefore, the trait modification observed entails a change of the normal distribution of the trait in the plants compared with the distribution observed in wild-type plants.

The term "transcript profile" refers to the expression levels of a set of genes in a cell in a particular state, particularly by comparison with the expression levels of that same set of genes in a cell of the same type in a reference state. For example, the transcript profile of a particular transcription factor in a suspension cell is the expression levels of a set of genes in a cell knocking out or overexpressing that transcription factor compared with the expression levels of that same set of genes in a suspension cell that has normal levels of that transcription factor. The transcript profile can be presented as a list of those genes whose expression level is significantly different between the two treatments, and the difference ratios. Differences and similarities between expression levels may also be evaluated and calculated using statistical and clustering methods.

"Ectopic expression or altered expression" in reference to a polynucleotide indicates that the pattern of expression in, e.g., a transgenic plant or plant tissue, is different from the expression pattern in a wild-type plant or a reference plant of the same species. The pattern of expression may also be compared with a reference expression pattern in a wild-type plant of the same species. For example, the polynucleotide or polypeptide is expressed in a cell or tissue type other than a cell or tissue type in which the sequence is expressed in the wild-type plant, or by expression at a time other than at the time the sequence is expressed in the wild-type plant, or by a response to different inducible agents, such as hormones or environmental signals, or at different expression levels (either higher or lower) compared with those found in a wild-type plant. The term also refers to altered expression patterns that are produced by lowering the levels of expression to below the detection level or completely abolishing

expression. The resulting expression pattern can be transient or stable, constitutive or inducible. In reference to a polypeptide, the term "ectopic expression or altered expression" further may relate to altered activity levels resulting from the interactions of the polypeptides with exogenous or endogenous modulators or from interactions with factors or as a result of the chemical modification of the polypeptides.

The term "overexpression" as used herein refers to a greater expression level of a gene in a plant, plant cell or plant tissue, compared to expression in a wild-type plant, cell or tissue, at any developmental or temporal stage for the gene. Overexpression can occur when, for example, the genes encoding one or more transcription factors are under the control of a strong expression signal, such as one of the promoters described herein (e.g., the cauliflower mosaic virus 35S transcription initiation region). Overexpression may occur throughout a plant or in specific tissues of the plant, depending on the promoter used, as described below.

Overexpression may take place in plant cells normally lacking expression of polypeptides functionally equivalent or identical to the present transcription factors. Overexpression may also occur in plant cells where endogenous expression of the present transcription factors or functionally equivalent molecules normally occurs, but such normal expression is at a lower level.. Overexpression thus results in a greater than normal production, or "overproduction" of the transcription factor in the plant, cell or tissue.

The term "transcription regulating region" refers to a DNA regulatory sequence that regulates expression of one or more genes in a plant when a transcription factor having one or more specific binding domains binds to the DNA regulatory sequence. Transcription factors of the present invention possess an AP2 domain, a B3 domain, or both of these binding domains. The AP2 domain of the transcription factor binds to a transcription regulating region comprising the motif CAACA, and the B3 domain of the same transcription factor binds to a transcription regulating region comprising the motif CACCTG. The transcription factors of the invention also comprise an amino acid subsequence that forms a transcription activation domain that regulates expression of one or more abiotic stress tolerance genes in a plant when the transcription factor binds to the regulating region.

The term "phase change" refers to a plant's progression from embryo to adult, and, by some definitions, the transition wherein flowering plants gain reproductive competency. It is believed that phase change occurs either after a certain number of cell divisions in the shoot apex of a developing plant, or when the shoot apex achieves a particular distance from the roots. Thus, altering the timing of phase changes may affect a plant's size, which, in turn, may affect yield and biomass.

A "sample" with respect to a material containing nucleic acid molecules may comprise a bodily fluid; an extract from a cell, chromosome, organelle, or membrane isolated from a cell; genomic DNA,

RNA, or cDNA in solution or bound to a substrate; a cell; a tissue; a tissue print; a forensic sample; and the like. In this context "substrate" refers to any rigid or semi-rigid support to which nucleic acid molecules or proteins are bound and includes membranes, filters, chips, slides, wafers, fibers, magnetic or nonmagnetic beads, gels, capillaries or other tubing, plates, polymers, and microparticles with a variety of surface forms including wells, trenches, pins, channels and pores. A substrate may also refer to a reactant in a chemical or biological reaction, or a substance acted upon (e.g., by an enzyme).

"Substantially purified" refers to nucleic acid molecules or proteins that are removed from their natural environment and are isolated or separated, and are at least about 60% free, preferably about 75% free, and most preferably about 90% free, from other components with which they are naturally associated.

DETAILED DESCRIPTION

Transcription Factors Modify Expression of Endogenous Genes

A transcription factor may include, but is not limited to, any polypeptide that can activate or repress transcription of a single gene or a number of genes. As one of ordinary skill in the art recognizes, transcription factors can be identified by the presence of a region or domain of structural similarity or identity to a specific consensus sequence or the presence of a specific consensus DNA-binding site or DNA-binding site motif (see, for example, Riechmann et al. (2000) *Science* 290: 2105-2110). The plant transcription factors may belong to the CAAT-element binding protein transcription factor family (Forsburg and Guarente (1989) *supra*).

Generally, the transcription factors encoded by the present sequences are involved in cell differentiation and proliferation and the regulation of growth. Accordingly, one skilled in the art would recognize that by expressing the present sequences in a plant, one may change the expression of autologous genes or induce the expression of introduced genes. By affecting the expression of similar autologous sequences in a plant that have the biological activity of the present sequences, or by introducing the present sequences into a plant, one may alter a plant's phenotype to one with improved traits related to osmotic stresses. The sequences of the invention may also be used to transform a plant and introduce desirable traits not found in the wild-type cultivar or strain. Plants may then be selected for those that produce the most desirable degree of over- or under-expression of target genes of interest and coincident trait improvement.

The sequences of the present invention may be from any species, particularly plant species, in a naturally occurring form or from any source whether natural, synthetic, semi-synthetic or recombinant. The sequences of the invention may also include fragments of the present amino acid

sequences. Where "amino acid sequence" is recited to refer to an amino acid sequence of a naturally occurring protein molecule, "amino acid sequence" and like terms are not meant to limit the amino acid sequence to the complete native amino acid sequence associated with the recited protein molecule.

In addition to methods for modifying a plant phenotype by employing one or more polynucleotides and polypeptides of the invention described herein, the polynucleotides and polypeptides of the invention have a variety of additional uses. These uses include their use in the recombinant production (i.e., expression) of proteins; as regulators of plant gene expression, as diagnostic probes for the presence of complementary or partially complementary nucleic acids (including for detection of natural coding nucleic acids); as substrates for further reactions, e.g., mutation reactions, PCR reactions, or the like; as substrates for cloning e.g., including digestion or ligation reactions; and for identifying exogenous or endogenous modulators of the transcription factors. In many instances, a polynucleotide comprises a nucleotide sequence encoding a polypeptide (or protein) or a domain or fragment thereof. Additionally, the polynucleotide may comprise a promoter, an intron, an enhancer region, a polyadenylation site, a translation initiation site, 5' or 3' untranslated regions, a reporter gene, a selectable marker, or the like. The polynucleotide can be single stranded or double stranded DNA or RNA. The polynucleotide optionally comprises modified bases or a modified backbone. The polynucleotide can be, e.g., genomic DNA or RNA, a transcript (such as an mRNA), a cDNA, a PCR product, a cloned DNA, a synthetic DNA or RNA, or the like. The polynucleotide can comprise a sequence in either sense or antisense orientations.

Expression of genes that encode transcription factors that modify expression of endogenous genes, polynucleotides, and proteins are well known in the art. In addition, transgenic plants comprising isolated polynucleotides encoding transcription factors may also modify expression of endogenous genes, polynucleotides, and proteins. Examples include Peng et al. (1997, *Genes Development* 11: 3194-3205) and Peng et al. (1999, *Nature*, 400: 256-261). In addition, many others have demonstrated that an *Arabidopsis* transcription factor expressed in an exogenous plant species elicits the same or very similar phenotypic response. See, for example, Fu et al. (2001, *Plant Cell* 13: 1791-1802); Nandi et al. (2000, *Curr. Biol.* 10: 215-218); Coupland (1995, *Nature* 377: 482-483); and Weigel and Nilsson (1995, *Nature* 377: 482-500).

In another example, Mandel et al. (1992, *Cell* 71-133-143) and Suzuki et al.(2001, *Plant J.* 28: 409-418) teach that a transcription factor expressed in another plant species elicits the same or very similar phenotypic response of the endogenous sequence, as often predicted in earlier studies of *Arabidopsis* transcription factors in *Arabidopsis* (see Mandel et al. 1992, *supra*; Suzuki et al. 2001, *supra*).

Other examples include Müller et al. (2001, *Plant J.* 28: 169-179); Kim et al. (2001, *Plant J.*

25: 247-259); Kyozuka and Shimamoto (2002, *Plant Cell Physiol.* 43: 130-135); Boss and Thomas (2002, *Nature*, 416: 847-850); He et al. (2000, *Transgenic Res.* 9: 223-227); and Robson et al. (2001, *Plant J.* 28: 619-631).

In yet another example, Gilmour et al. (1998, *Plant J.* 16: 433-442) teach an *Arabidopsis* AP2 transcription factor, CBF1 (SEQ ID NO: 96), which, when overexpressed in transgenic plants, increases plant freezing tolerance. Jaglo et al. ((2001) *Plant Physiol.* 127: 910-917) further identified sequences in *Brassica napus* which encode CBF-like genes and that transcripts for these genes accumulated rapidly in response to low temperature. Transcripts encoding CBF-like proteins were also found to accumulate rapidly in response to low temperature in wheat, as well as in tomato. An alignment of the CBF proteins from *Arabidopsis*, *B. napus*, wheat, rye, and tomato revealed the presence of conserved consecutive amino acid residues, PKK/RPAGRxKFxETRHP and DSAWR, that bracket the AP2/EREBP DNA binding domains of the proteins and distinguish them from other members of the AP2/EREBP protein family. (See Jaglo et al. *supra.*)

Transcription factors mediate cellular responses and control traits through altered expression of genes containing cis-acting nucleotide sequences that are targets of the introduced transcription factor. It is well appreciated in the Art that the effect of a transcription factor on cellular responses or a cellular trait is determined by the particular genes whose expression is either directly or indirectly (e.g., by a cascade of transcription factor binding events and transcriptional changes) altered by transcription factor binding. In a global analysis of transcription comparing a standard condition with one in which a transcription factor is overexpressed, the resulting transcript profile associated with transcription factor overexpression is related to the trait or cellular process controlled by that transcription factor. For example, the PAP2 gene (and other genes in the MYB family) have been shown to control anthocyanin biosynthesis through regulation of the expression of genes known to be involved in the anthocyanin biosynthetic pathway (Bruce et al. (2000) *Plant Cell* 12: 65-79; and Borevitz et al. (2000) *Plant Cell* 12: 2383-2393). Further, global transcript profiles have been used successfully as diagnostic tools for specific cellular states (e.g., cancerous vs. non-cancerous; Bhattacharjee et al. (2001) *Proc. Natl. Acad. Sci.* USA 98: 13790-13795; and Xu et al. (2001) *Proc Natl Acad Sci*, USA 98: 15089-15094). Consequently, it is evident to one skilled in the art that similarity of transcript profile upon overexpression of different transcription factors would indicate similarity of transcription factor function.

CCAAT-element binding protein transcription factor family

The CAAT family of transcription factors, also be referred to as the "CCAAT" or "CCAAT-box" family, are characterized by their ability to bind to the CCAAT-box element located 80 to 300 bp

5' from a transcription start site (Gelinas et al. (1985) *Nature* 313: 323-325). The CCAAT-box is a conserved cis-acting regulatory element with the consensus sequence CCAAT that is found in the promoters of genes from all eukaryotic species. The element can act in either orientation, alone or as multimeric regions with possible cooperation with other cis regulatory elements (Tasanen et al. (1992) (*J. Biol. Chem.* 267: 11513-11519). It has been estimated that 25% of eukaryotic promoters harbor this element (Bucher (1988) *J. Biomol. Struct. Dyn.* 5: 1231-1236). CCAAT-box elements have been shown to function in the regulation of gene expression in plants (Rieping and Schoffl (1992) *Mol. Gen. Genet.* 231: 226-232; Kehoe et al. (1994) *Plant Cell* 6: 1123-1134; Ito et al. (1995) *Plant Cell Physiol.* 36: 1281-1289). Several reports have described the importance of the CCAAT-binding element for regulated expression; including the regulation of genes that are responsive to light (Kusnetsov et al. (1999) *J. Biol. Chem.* 274: 36009-36014; Carre and Kay (1995) *Plant Cell* 7: 2039-2051) as well as stress (Rieping and Schoffl (1992) *supra*). Specifically, a CCAAT-box motif was shown to be important for the light regulated expression of the CAB2 promoter in *Arabidopsis*, however, the proteins that bind to the site were not identified (Carre and Kay (1995) *supra*). To date, no specific *Arabidopsis* CCAAT-box binding protein has been functionally associated with its corresponding target genes. In October of 2002 at an EPSO meeting on Plant Networks, a seminar was given by Detlef Weigel (Tuebingen) on the control of the AGAMOUS (a floral organ identity gene) gene in *Arabidopsis*. In order to find important cis-elements that regulate AGAMOUS activity, he aligned the promoter regions from 29 different Brassicaceae species and showed that there were two highly conserved regions; one well characterized site that binds LEAFY/WUS heterodimers and another putative CCAAT-box binding motif. We have discovered several CCAAT-box genes that regulate flowering time and are candidates for binding to the AGAMOUS promoter. One of these genes, G485, is a HAP3-like protein that is closely related to G481. Gain of function and loss of function studies on G485 reveal opposing effects on flowering time, indicating that the gene is both sufficient to act as a floral activator, and is also necessary in that role within the plant.

The first proteins identified that bind to the CCAAT-box element were identified in yeast. The CCAAT-box transcription factors bind as hetero-tetrameric complex called the HAP complex (heme activator protein complex) or the CCAAT binding factor (Forsburg and Guarente (1988) *Mol. Cell Biol.* 8: 647-654). The HAP complex in yeast is composed of at least four subunits, HAP2, HAP3, HAP4 and HAP5. In addition, the proteins that make up the HAP2,3,4,5 complex are represented by single genes. Their function is specific for the activation of genes involved in mitochondrial biogenesis and energy metabolism (Dang et al. (1996) *Mol. Microbiol.* 22:681-692). In mammals, the CCAAT binding factor is a trimeric complex consisting of NF-YA (HAP2-like), NF-YB (HAP3-like) and NF-YC (HAP5-like) subunits (Maity and de Crombrugghe (1998) *Trends Biochem. Sci.* 23: 174-178). In

plants, analogous members of the CCAAT binding factor complex are represented by small gene families, and it is likely that these genes play a more complex role in regulating gene transcription. In *Arabidopsis* there are ten members of the HAP2 subfamily, ten members of the HAP3 subfamily, thirteen members of the HAP5 subfamily. Plants and mammals, however, do not appear to have a protein equivalent of HAP4 of yeast. HAP4 is not required for DNA binding in yeast although it provides the primary activation domain for the complex (McNabb et al. (1995) *Genes Dev.* 9: 47-58; Olesen and Guarente (1990) *Genes Dev.* 4, 1714-1729).

In mammals, the CCAAT-box element is found in the promoters of many genes and it is therefore been proposed that CCAAT binding factors serve as general transcriptional regulators that influence the frequency of transcriptional initiation (Maity and de Crombrugghe (1998) *supra*). CCAAT binding factors, however, can serve to regulate target promoters in response to environmental cues and it has been demonstrated that assembly of CCAAT binding factors on target promoters occurs in response to a variety signals (Myers et al. (1986) Myers et al. (1986) *Science* 232: 613-618; Maity and de Crombrugghe (1998) *supra*; Bezhani et al. (2001) *J. Biol. Chem.* 276: 23785-23789). Mammalian CP1 and NF-Y are both heterotrimeric CCAAT binding factor complexes (Johnson and McKnight (1989) *Ann. Rev. Biochem.* 58: 799-839. Plant CCAAT binding factors are assumed to be trimeric, as is the case in mammals, however, they could associate with other transcription factors on target promoters as part of a larger complex. The CCAAT box is generally found in close proximity of other promoter elements and it is generally accepted that the CCAAT binding factor functions synergistically with other transcription factors in the regulation of transcription. In addition, it has recently been shown that a HAP3-like protein from rice, OsNF-YB1, interacts with a MADS-box protein OsMADS18 in vitro (Masiero et al. (2002) *J. Biol. Chem.* 277: 26429-26435). It was also shown that the in vitro ternary complex between these two types of transcription factors requires that both; OsNF-YB1 form a dimer with a HAP5-like protein, and that OsMADS18 form a heterodimer with another MADS-box protein. Interestingly, the OsNF-YB1/HAP5 protein dimer is incapable of interacting with HAP2-like subunits and therefore cannot bind the CCAAT element. The authors therefore speculate that there is a select set of HAP3-like proteins in plants that act on non-CCAAT promoter elements by virtue of their interaction with other non-CCAAT transcription factors (Masiero et al. (2002) *supra*). In support of this, HAP3/HAP5 subunit dimers have been shown to be able to interact with TFIID in the absence of HAP2 subunits (Romier et al. (2003) *J. Biol. Chem.* 278: 1336-1345).

The CCAAT-box motif is found in the promoters of a variety of plant genes. In addition, the expression pattern of many of the HAP-like genes in *Arabidopsis* shows developmental regulation. We have used RT-PCR to analyze the endogenous expression of 31 of the 34 CCAAT-box proteins . Our

findings suggest that while most of the CCAAT-box gene transcripts are found ubiquitously throughout the plant, in more than half of the cases, the genes are predominantly expressed in flower, embryo and/or silique tissues. Cell-type specific localization of the CCAAT genes in *Arabidopsis* would be very informative and could help determine the activity of various CCAAT genes in the plant.

Genetic analysis has determined the function of one *Arabidopsis* CCAAT gene, LEAFY COTYLEDON (LEC1). LEC1 is a HAP3 subunit homolog that accumulates only during seed development. *Arabidopsis* plants carrying a mutation in the LEC1 gene display embryos that are intolerant to desiccation and that show defects in seed maturation (Lotan et al. (1998) *Cell* 93: 1195-1205). This phenotype can be rescued if the embryos are allowed to grow before the desiccation process occurs during normal seed maturation. This result suggests LEC1 has a role in allowing the embryo to survive desiccation during seed maturation. The mutant plants also possess trichomes, or epidermal hairs on their cotyledons, a characteristic that is normally restricted to adult tissues like leaves and stems. Such an effect suggests that LEC1 also plays a role in specifying embryonic organ identity. In addition to the mutant analysis, the ectopic expression (unregulated overexpression) of the wild type LEC1 gene induces embryonic programs and embryo development in vegetative cells consistent with its role in coordinating higher plant embryo development. The ortholog of LEC1 has been identified recently in maize. The expression pattern of ZmLEC1 in maize during somatic embryo development is similar to that of LEC1 in *Arabidopsis* during zygotic embryo development (Zhang et al. (2002) *Planta* 215:191-194).

Matching the CCAAT transcription factors with target promoters and the analysis of the knockout and overexpression mutant phenotypes will help sort out whether these proteins act specifically or non-specifically in the control of plant pathways. The fact that CCAAT-box elements are not present in most plant promoters suggests that plant CCAAT binding factors most likely do not function as general components of the transcriptional machinery. In addition, the very specific role of the LEC1 protein in plant developmental processes supports the idea that CCAAT-box binding complexes play very specific roles in plant growth and development.

The domain structure of CCAAT-element binding transcription factors and novel conserved domains in *Arabidopsis* and other species

Plant CCAAT binding factors potentially bind DNA as heterotrimers composed of HAP2-like, HAP3-like and HAP5-like subunits. All subunits contain regions that are required for DNA binding and subunit association. The subunit proteins appear to lack activation domains; therefore, that function must come from proteins with which they interact on target promoters. No proteins that provide the activation domain function for CCAAT binding factors have been identified in plants. In

yeast, however, the HAP4 protein provides the primary activation domain (McNabb et al. (1995) *Genes Dev.* 9: 47-58; Olesen and Guarente (1990) *Genes Dev.* 4, 1714-1729).

HAP2-, HAP3- and HAP5-like proteins have two highly conserved sub-domains, one that functions in subunit interaction and the other that acts in a direct association with DNA. Outside these two regions, non-paralogous *Arabidopsis* HAP-like proteins are quite divergent in sequence and in overall length.

The general domain structure of HAP3 proteins is found in Figure 5. HAP3 proteins contain an amino-terminal A domain, a central B domain and a carboxy-terminal C domain. There is very little sequence similarity between HAP3 proteins in the A and C domains; it is therefore reasonable to assume that the A and C domains could provide a degree of functional specificity to each member of the HAP3 subfamily. The B domain is the conserved region that specifies DNA binding and subunit association.

In Figures 6A-6F, HAP3 proteins from *Arabidopsis*, soybean, rice and corn are aligned with G481, with the A, B and C domains and the DNA binding and subunit interaction domains indicated. As can be seen in Figure 6B-6C, the B domain of the non-LEC1-like clade (identified in Figures 3 and 4) may be distinguished by the amino acid residues:

Ser/Gly-Arg-Ile/Leu-Met-Lys-(Xaa)$_2$-Lys/Ile/Val-Pro-Xaa-**Asn**-Ala/Gly-Lys-Ile/Val-Ser/Ala/Gly-**Lys**-Asp/Glu-Ala/Ser-Lys-Glu/Asp/Gln-Thr/Ile-Xaa-Gln-Glu-Cys-Val/Ala-Ser/Thr-Glu-Phe-Ile-Ser-Phe-Ile/Val/His-Thr/Ser-[Pro]-Gly/Ser/Cys-Glu-Ala/Leu-Ser/Ala-Asp/Glu/Gly-Lys/Glu-Cys-Gln/His-Arg/Lys-Glu-Lys/Asn-Arg-Lys-Thr-Ile/Val-**Asn-Gly**-Asp/Glu-Asp-Leu/Ile-Xaa-Trp/Phe-Ala-Met/Ile/Leu-Xaa-Thr/Asn-Leu-Gly-Phe/Leu-Glu/Asp-Xaa-Tyr-(Xaa)$_2$-Pro/Gln/Ala-Leu/Val-Lys/Gly;

where Xaa can be any amino acid. The proline residue that appears in brackets is an additional residue that was found in only one sequence (not shown in Figure 6B). The boldfaced residues that appear here and in the consensus sequences of Figures 6B - 6C in their present positions are uniquely found in the non-LEC1-like clade, and may be used to identify members of this clade. The G482-like subclade may be delineated by the underlined serine residue in its present position here and in the consensus sequence of Figures 6B-6C. More generally, the non-LEC1-like clade is distinguished by a B domain comprising:

Asn-(Xaa)$_4$-Lys-(Xaa)$_{33-34}$-Asn-Gly;

and the G482 subclade is distinguished by a B-domain comprising:

Ser-$(Xaa)_9$-Asn-$(Xaa)_4$-Lys-$(Xaa)_{33-34}$-Asn-Gly.

Overexpression of these polypeptides confers increased abiotic stress tolerance in a transgenic plant, as compared to a non-transformed plant that does not overexpress the polypeptide.

Table 1 shows the polypeptides identified by SEQ ID NO; Mendel Gene ID (GID) No.; the transcription factor family to which the polypeptide belongs, and conserved B domains of the polypeptide. The first column shows the polypeptide SEQ ID NO; the second column the species and identifier (GID, GenBank accession no., or other identifier); the third column shows the conserved domain in amino acid coordinates; the fourth column shows the B domain; and the fifth column shows the percentage identity to G482. The sequences are arranged in descending order of percentage identity to G482.

Table 1. Gene families and B domains

Polypeptide SEQ ID NO:	Species/ GID No., Accession No., or Identifier	CCAAT-box binding conserved domain in Amino Acid Coordinates	B Domain	% ID to CCAAT-box binding conserved domain of G482
4	At/G482	26-116	REQDRFLPIANVSRIMKKALPANAKISKDAKET MQECVSEFISFVTGEASDKCQKEKRKTINGDD LLWAMTTLGFEDYVEPLKVYLQRFRE	100%
20	Gm/G3475	23-113	REQDRFLPIANVSRIMKKALPANAKISKDAKET VQECVSEFISFITGEASDKCQREKRKTINGDDL LWAMTTLGFEDYVEPLKGYLQRFRE	95%
86	Gm/3478	23-113	REQDRFLPIANVSRIMKKALPANAKISKDAKET VQECVSEFISFITGEASDKCQREKRKTINGDDL LWAMTTLGFEDYVEPLKGYLQRFRE	95%
6	At/G485	20-110	REQDRFLPIANVSRIMKKALPANAKISKDAKET VQECVSEFISFITGEASDKCQREKRKTINGDDL LWAMTTLGFEDYVEPLKVYLQKYRE	94%
18	Gm/G3476	26-116	REQDRFLPIANVSRIMKKALPANAKISKDAKET VQECVSEFISFITGEASDKCQREKRKTINGDDL LWAMTTLGFEEYVEPLKIYLQRFRE	94%
48	Zm/ CLUSTER 90408_1	22-112	REQDRFLPIANVSRIMKKALPANAKISKDAKET VQECVSEFISFITGEASDKCQREKRKTINGDDL LWAMTTLGFEDYVEPLKHYLHKFRE	93%
48	Zm/G3435	22-112	REQDRFLPIANVSRIMKKALPANAKISKDAKET VQECVSEFISFITGEASDKCQREKRKTINGDDL LWAMTTLGFEDYVEPLKHYLHKFRE	93%
50	Zm/G3436 CLUSTER 90408_2	20-110	REQDRFLPIANVSRIMKKALPANAKISKDAKET VQECVSEFISFITGEASDKCQREKRKTINGDDL LWAMTTLGFEDYVEPLKLYLHKFRE	93%
92	Os/G3397 AC120529	23-113	REQDRFLPIANVSRIMKKALPANAKISKDAKET VQECVSEFISFITGEASDKCQREKRKTINGDDL LWAMTTLGFEDYVDPLKHYLHKFRE	92%
80	Gm/G3472	25-115	REQDRFLPIANVSRIMKKALPANAKISKEAKET VQECVSEFISFITGEASDKCQKEKRKTINGDDL LWAMTTLGFEEYVEPLKVYLHKYRE	92%
82	Gm/G3474 CLUSTER 33504_1	25-115	REQDRFLPIANVSRIMKKALPANAKISKEAKET VQECVSEFISFITGEASDKCQKEKRKTINGDDL LWAMTTLGFEDYVDPLKIYLHKYRE	91%
76	Os/G3398 AP005193	21-111	REQDRFLPIANVSRIMKRALPANAKISKDAKET VQECVSEFISFITGEASDKCQREKRKTINGDDL LWAMTTLGFEDYIDPLKLYLHKFRE	90%
94	Zm/G3437	54-144	KEQDRFLPIANVSRIMKRSLPANAKISKEAKET VQECVSEFISFVTGEASDKCQREKRKTINGDDL LWAMTTLGFEAYVAPLKSYLNRYRE	87%
28	Os/CLUSTER26105_1	38-127	VRQDRFLPIANISRIMKKAIPANGKIAKDAKET VQECVSEFISFITSEASDKCQREKRKTINGDDLL WAMATLGFEDYIEPLKVYLQKYRE	86%
78	Zm/G3434	18-108	REQDRFLPIANISRIMKKAVPANGKIAKDAKET LQECVSEFISFVTSEASDKCQKEKRKTINGDDL LWAMATLGFEEYVEPLKIYLQKYKE	86%
31	Os/ OSC30077	57-147	KEQDRFLPIANVSRIMKRSLPANAKISKESKET VQECVSEFISFVTGEASDKCQREKRKTINGDDL	86%

			LWAMTTLGFEAYVGPLKSYLNRYRE	
88	Os/G3394	37-127	VRQDRFLPIANISRIMKKAIPANGKIAKDAKET VQECVSEFISFITSEASDKCQREKRKTINGDDLL WAMATLGFEDYIEPLKVYLQKYRE	86%
24	Gm/G3471	26-116	REQDRYLPIANISRIMKKALPPNGKIAKDAKDT MQECVSEFISFITSEASEKCQKEKRKTINGDDL LWAMATLGFEDYIEPLKVYLARYRE	85%
26	Gm/G3470 CLUSTER 4778_3	27-117	REQDRYLPIANISRIMKKALPPNGKIAKDAKDT MQECVSEFISFITSEASEKCQKEKRKTINGDDL LWAMATLGFEDYIEPLKVYLARYRE	85%
52	Gm/G3473	23-114	REQDRFLPIANVSRIMKKALPANAKISKDAKET VQECVSEFISFHSPGGLAGECQKEKRKTINGDD LLWAMTTLGFEEYVEPLKVYLHKYRE	85%
8	At/G1364	29-119	REQDRFLPIANISRIMKRGLPANGKIAKDAKEI VQECVSEFISFVTSEASDKCQREKRKTINGDDL LWAMATLGFEDYMEPLKVYLMRYRE	85%
10	At/G2345	28-118	REQDRFLPIANISRIMKRGLPLNGKIAKDAKET MQECVSEFISFVTSEASDKCQREKRKTINGDDL LWAMATLGFEDYIDPLKVYLMRYRE	85%
86	Gm/G3477	27-117	REQDRYLPIANISRIMKKALPPNGKIAKDAKDT MQECVSEFISFITSEASEKCQKEKRKTINGDDL LWAMATLGFEDYIEPLKVYLARYRE	85%
2	At/G481	20-110	REQDRYLPIANISRIMKKALPPNGKIGKDAKDT VQECVSEFISFITSEASDKCQKEKRKTVNGDDL LWAMATLGFEDYLEPLKIYLARYRE	83%
72	At/ G1781	35-125	KEQDRFLPIANVGRIMKKVLPGNGKISKDAKE TVQECVSEFISFVTGEASDKCQREKRKTINGDD IIWAITTLGFEDYVAPLKVYLCKYRD	83%
74	Os/G3395	19-109	REQDRFLPIANISRIMKKAVPANGKIAKDAKET LQECVSEFISFVTSEASDKCQKEKRKTINGEDL LFAMGTLGFEEYVDPLKIYLHKYRE	83%
	Os/ AP004366	19-109	REQDRFLPIANISRIMKKAVPANGKIAKDAKET LQECVSEFISFVTSEASDKCQKEKRKTINGEDL LFAMGTLGFEEYVDPLKIYLHKYRE	83%
70	At/G1248	50-140	KEQDRLLPIANVGRIMKNILPANAKVSKEAKE TMQECVSEFISFVTGEASDKCHKEKRKTVNGD DICWAMANLGFDDYAAQLKKYLHRYRV	77%
90	Os/G3396	21-111	KEQDRFLPIANIGRIMRRAVPENGKIAKDSKES VQECVSEFISFITSEASDKCLKEKRKTINGDDLI WSMGTLGFEDYVEPLKLYLRLYRE	75%
60	At/G1821 L1L	28-118	REQDRFMPIANVIRIMRRILPAHAKISDDSKETI QECVSEYISFITGEANERCQREQRKTITAEDVL WAMSKLGFDDYIEPLTLYLHRYRE	69%
	At/ AAC39488 LEC1	28-118	REQDQYMPIANVIRIMRKTLPSHAKISDDAKET IQECVSEYISFVTGEANERCQREQRKTITAEDIL WAMSKLGFDNYVDPLTVFINRYRE	67%
	At/G486	2-92	TDEDRLLPIANVGRLMKQILPSNAKISKEAKQT VQECATEFISFVTCEASEKCHRENRKTVNGDDI WWALSTLGLDNYADAVGRHLHKYRE	60%

Abbreviations: At Arabidopsis thaliana
Gm Glycine max
Os Oryza sativa
Zm Zea mays

The transcription factors of the present invention each possess a B or conserved domain, including the orthologs of G482 found by BLAST analysis, as described below. Generally, the B domain of the transcription factors will bind to a transcription-regulating region comprising the motif CCAAT. As shown in Table 1, the B domains of G481, G485 and rice G3395 are at least 83% identical to the corresponding domains of G482, and all four of these transcription factors, which rely on the binding specificity of their B domains, have similar or identical functions in plants, conferring increased abiotic, including osmotic, stress tolerance when overexpressed.

Polypeptides and Polynucleotides of the Invention

The present invention provides, among other things, transcription factors (TFs), and transcription factor homolog polypeptides, and isolated or recombinant polynucleotides encoding the polypeptides, or novel sequence variant polypeptides or polynucleotides encoding novel variants of transcription factors derived from the specific sequences provided here. These polypeptides and polynucleotides may be employed to modify a plant's characteristics.

Exemplary polynucleotides encoding the polypeptides of the invention were identified in the *Arabidopsis thaliana* GenBank database using publicly available sequence analysis programs and parameters. Sequences initially identified were then further characterized to identify sequences comprising specified sequence strings corresponding to sequence motifs present in families of known transcription factors. In addition, further exemplary polynucleotides encoding the polypeptides of the invention were identified in the plant GenBank database using publicly available sequence analysis programs and parameters. Sequences initially identified were then further characterized to identify sequences comprising specified sequence strings corresponding to sequence motifs present in families of known transcription factors. Polynucleotide sequences meeting such criteria were confirmed as transcription factors.

Additional polynucleotides of the invention were identified by screening *Arabidopsis thaliana* and/or other plant cDNA libraries with probes corresponding to known transcription factors under low stringency hybridization conditions. Additional sequences, including full length coding sequences were subsequently recovered by the rapid amplification of cDNA ends (RACE) procedure, using a commercially available kit according to the manufacturer's instructions. Where necessary, multiple rounds of RACE are performed to isolate 5' and 3' ends. The full-length cDNA was then recovered by a routine end-to-end polymerase chain reaction (PCR) using primers specific to the isolated 5' and 3' ends. Exemplary sequences are provided in the Sequence Listing.

The polynucleotides of the invention can be or were ectopically expressed in overexpressor or

knockout plants and the changes in the characteristic(s) or trait(s) of the plants observed. Therefore, the polynucleotides and polypeptides can be employed to improve the characteristics of plants.

The polynucleotides of the invention can be or were ectopically expressed in overexpressor plant cells and the changes in the expression levels of a number of genes, polynucleotides, and/or proteins of the plant cells observed. Therefore, the polynucleotides and polypeptides can be employed to change expression levels of a genes, polynucleotides, and/or proteins of plants.

The CCAAT family members under study

The correct sequences for G482, and trait disclosures for G481, G482 and G485, were first disclosed in U.S. Provisional Patent Application 60/166,228, filed November 17, 1999.

G481, G482 and G485 (polynucleotide SEQ ID NOs: 1, 3 and 5) were chosen for study based on observations that *Arabidopsis* plants overexpressing these genes had resistance to abiotic stresses, such as osmotic stress, and including drought-related stress (see Example VIII, below). G481, G482 and G485 are members of the CCAAT family, proteins that act in a multi-subunit complex and are believed to bind CCAAT boxes in promoters of target genes as trimers or tetramers.

In *Arabidopsis*, three types of CCAAT binding proteins exist: HAP2, HAP3 and HAP5. The G481, G482 and G485 polypeptides, as well as a number of other proteins in the *Arabidopsis* proteome, belong to the HAP3 class. As reported in the scientific literature thus far, only two genes from the HAP3 class have been functionally analyzed to a substantial degree. These are LEAFY COTYLEDON1 (LEC1) and its most closely related subunit, LEC1-LIKE (L1L). LEC1 and L1L are expressed primarily during seed development. Both appear to be essential for embryo survival of desiccation during seed maturation (Kwong et al. (2003) *Plant Cell* 15: 5-18). LEC1 is a critical regulator required for normal development during the early and late phases of embryogenesis that is sufficient to induce embryonic development in vegetative cells. Kwong et al. showed that ten *Arabidopsis* HAP3 subunits can be divided into two classes based on sequence identity in their central, conserved B domain. LEC1 and L1L constitute LEC1-type HAP3 subunits, whereas the remaining HAP3 subunits were designated non-LEC1-type.

Phylogenetic trees based on sequential relatedness of the HAP3 genes are shown in Figure 3 and 4. As can be seen in these figures, G1364 and G2345 are closely related to G481, and G482 and G485 are more related to G481 than either LEC1 or L1L, which are found on somewhat more distant nodes.

Producing Polypeptides

The polynucleotides of the invention include sequences that encode transcription factors and transcription factor homolog polypeptides and sequences complementary thereto, as well as unique fragments of coding sequence, or sequence complementary thereto. Such polynucleotides can be, e.g., DNA or RNA, e.g., mRNA, cRNA, synthetic RNA, genomic DNA, cDNA synthetic DNA, oligonucleotides, etc. The polynucleotides are either double-stranded or single-stranded, and include either, or both sense (i.e., coding) sequences and antisense (i.e., non-coding, complementary) sequences. The polynucleotides include the coding sequence of a transcription factor, or transcription factor homolog polypeptide, in isolation, in combination with additional coding sequences (e.g., a purification tag, a localization signal, as a fusion-protein, as a pre-protein, or the like), in combination with non-coding sequences (e.g., introns or inteins, regulatory elements such as promoters, enhancers, terminators, and the like), and/or in a vector or host environment in which the polynucleotide encoding a transcription factor or transcription factor homolog polypeptide is an endogenous or exogenous gene.

A variety of methods exist for producing the polynucleotides of the invention. Procedures for identifying and isolating DNA clones are well known to those of skill in the art, and are described in, e.g., Berger and Kimmel, Guide to Molecular Cloning Techniques, *Methods in Enzymology,* vol. 152 Academic Press, Inc., San Diego, CA ("Berger"); Sambrook et al. Molecular Cloning - A Laboratory Manual (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 ("Sambrook") and Current Protocols in Molecular Biology, Ausubel et al. eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 2000) ("Ausubel").

Alternatively, polynucleotides of the invention, can be produced by a variety of in vitro amplification methods adapted to the present invention by appropriate selection of specific or degenerate primers. Examples of protocols sufficient to direct persons of skill through in vitro amplification methods, including the polymerase chain reaction (PCR) the ligase chain reaction (LCR), Qbeta-replicase amplification and other RNA polymerase mediated techniques (e.g., NASBA), e.g., for the production of the homologous nucleic acids of the invention are found in Berger (supra), Sambrook (supra), and Ausubel (supra), as well as Mullis et al. (1987) PCR Protocols A Guide to Methods and Applications (Innis et al. eds) Academic Press Inc. San Diego, CA (1990) (Innis). Improved methods for cloning in vitro amplified nucleic acids are described in Wallace et al. US Pat. No. 5,426,039. Improved methods for amplifying large nucleic acids by PCR are summarized in Cheng et al. (1994) *Nature* 369: 684-685 and the references cited therein, in which PCR amplicons of up to 40kb are generated. One of skill will appreciate that essentially any RNA can be converted into a double stranded DNA suitable for restriction digestion, PCR expansion and sequencing using reverse

transcriptase and a polymerase. See, e.g., Ausubel, Sambrook and Berger, all *supra*.

Alternatively, polynucleotides and oligonucleotides of the invention can be assembled from fragments produced by solid-phase synthesis methods. Typically, fragments of up to approximately 100 bases are individually synthesized and then enzymatically or chemically ligated to produce a desired sequence, e.g., a polynucleotide encoding all or part of a transcription factor. For example, chemical synthesis using the phosphoramidite method is described, e.g., by Beaucage et al. (1981) *Tetrahedron Letters* 22: 1859-1869; and Matthes et al. (1984) *EMBO J.* 3: 801-805. According to such methods, oligonucleotides are synthesized, purified, annealed to their complementary strand, ligated and then optionally cloned into suitable vectors. And if so desired, the polynucleotides and polypeptides of the invention can be custom ordered from any of a number of commercial suppliers.

Homologous Sequences

Sequences homologous, i.e., that share significant sequence identity or similarity, to those provided in the Sequence Listing , derived from *Arabidopsis thaliana* or from other plants of choice, are also an aspect of the invention. Homologous sequences can be derived from any plant including monocots and dicots and in particular agriculturally important plant species, including but not limited to, crops such as soybean, wheat, corn (maize), potato, cotton, rice, rape, oilseed rape (including canola), sunflower, alfalfa, clover, sugarcane, and turf; or fruits and vegetables, such as banana, blackberry, blueberry, strawberry, and raspberry, cantaloupe, carrot, cauliflower, coffee, cucumber, eggplant, grapes, honeydew, lettuce, mango, melon, onion, papaya, peas, peppers, pineapple, pumpkin, spinach, squash, sweet corn, tobacco, tomato, tomatillo, watermelon, rosaceous fruits (such as apple, peach, pear, cherry and plum) and vegetable brassicas (such as broccoli, cabbage, cauliflower, Brussels sprouts, and kohlrabi). Other crops, including fruits and vegetables, whose phenotype can be changed and which comprise homologous sequences include barley; rye; millet; sorghum; currant; avocado; citrus fruits such as oranges, lemons, grapefruit and tangerines, artichoke, cherries; nuts such as the walnut and peanut; endive; leek; roots such as arrowroot, beet, cassava, turnip, radish, yam, and sweet potato; and beans. The homologous sequences may also be derived from woody species, such pine, poplar and eucalyptus, or mint or other labiates. In addition, homologous sequences may be derived from plants that are evolutionarily related to crop plants, but which may not have yet been used as crop plants. Examples include deadly nightshade (*Atropa belladona*), related to tomato; jimson weed (Datura strommium), related to peyote; and teosinte (*Zea* species), related to corn (maize).

Orthologs and Paralogs

Homologous sequences as described above can comprise orthologous or paralogous sequences. Several different methods are known by those of skill in the art for identifying and defining these functionally homologous sequences. Three general methods for defining orthologs and paralogs are described; an ortholog, paralog or homolog may be identified by one or more of the methods described below.

Orthologs and paralogs are evolutionarily related genes that have similar sequence and similar functions. Orthologs are structurally related genes in different species that are derived by a speciation event. Paralogs are structurally related genes within a single species that are derived by a duplication event.

Within a single plant species, gene duplication may cause two copies of a particular gene, giving rise to two or more genes with similar sequence and often similar function known as paralogs. A paralog is therefore a similar gene formed by duplication within the same species. Paralogs typically cluster together or in the same clade (a group of similar genes) when a gene family phylogeny is analyzed using programs such as CLUSTAL (Thompson et al. (1994) *Nucleic Acids Res.* 22: 4673-4680; Higgins et al. (1996) *Methods Enzymol.* 266: 383-402). Groups of similar genes can also be identified with pair-wise BLAST analysis (Feng and Doolittle (1987) *J. Mol. Evol.* 25: 351-360). For example, a clade of very similar MADS domain transcription factors from *Arabidopsis* all share a common function in flowering time (Ratcliffe et al. (2001) *Plant Physiol.* 126: 122-132), and a group of very similar AP2 domain transcription factors from *Arabidopsis* are involved in tolerance of plants to freezing (Gilmour et al. (1998) *Plant J.* 16: 433-442). Analysis of groups of similar genes with similar function that fall within one clade can yield sub-sequences that are particular to the clade. These sub-sequences, known as consensus sequences, can not only be used to define the sequences within each clade, but define the functions of these genes; genes within a clade may contain paralogous sequences, or orthologous sequences that share the same function (see also, for example, Mount (2001), in Bioinformatics: Sequence and Genome Analysis Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, page 543.)

Speciation, the production of new species from a parental species, can also give rise to two or more genes with similar sequence and similar function. These genes, termed orthologs, often have an identical function within their host plants and are often interchangeable between species without losing function. Because plants have common ancestors, many genes in any plant species will have a corresponding orthologous gene in another plant species. Once a phylogenic tree for a gene family of one species has been constructed using a program such as CLUSTAL (Thompson et al. (1994) Nucleic Acids Res. 22: 4673-4680; Higgins et al. (1996) *supra*) potential orthologous sequences can be placed

into the phylogenetic tree and their relationship to genes from the species of interest can be determined. Orthologous sequences can also be identified by a reciprocal BLAST strategy. Once an orthologous sequence has been identified, the function of the ortholog can be deduced from the identified function of the reference sequence.

Transcription factor gene sequences are conserved across diverse eukaryotic species lines (Goodrich et al. (1993) *Cell* 75: 519-530; Lin et al. (1991) *Nature* 353: 569-571; Sadowski et al. (1988) *Nature* 335: 563-564). Plants are no exception to this observation; diverse plant species possess transcription factors that have similar sequences and functions.

Orthologous genes from different organisms have highly conserved functions, and very often essentially identical functions (Lee et al. (2002) *Genome Res.* 12: 493-502; Remm et al. (2001) *J. Mol. Biol.* 314: 1041-1052). Paralogous genes, which have diverged through gene duplication, may retain similar functions of the encoded proteins. In such cases, paralogs can be used interchangeably with respect to certain embodiments of the instant invention (for example, transgenic expression of a coding sequence). An example of such highly related paralogs is the CBF family, with three well-defined members in *Arabidopsis* and at least one ortholog in *Brassica napus* (SEQ ID NOs: 96, 98, 100, and 102, respectively), all of which control pathways involved in both freezing and drought stress (Gilmour et al. (1998) *Plant J.* 16: 433-442; Jaglo et al. (1998) *Plant Physiol.* 127: 910-917).

The following references represent a small sampling of the many studies that demonstrate that conserved transcription factor genes from diverse species are likely to function similarly (i.e., regulate similar target sequences and control the same traits), and that transcription factors may be transformed into diverse species to confer or improve traits.

(1) The *Arabidopsis* NPR1 gene regulates systemic acquired resistance (SAR); over-expression of NPR1 leads to enhanced resistance in *Arabidopsis*. When either *Arabidopsis* NPR1 or the rice NPR1 ortholog was overexpressed in rice (which, as a monocot, is diverse from *Arabidopsis*), challenge with the rice bacterial blight pathogen *Xanthomonas oryzae* pv. Oryzae, the transgenic plants displayed enhanced resistance (Chern et al. (2001) *Plant J.* 27: 101-113). NPR1 acts through activation of expression of transcription factor genes, such as TGA2 (Fan and Dong (2002) *Plant Cell* 14: 1377-1389).

(2) E2F genes are involved in transcription of plant genes for proliferating cell nuclear antigen (PCNA). Plant E2Fs share a high degree of similarity in amino acid sequence between monocots and dicots, and are even similar to the conserved domains of the animal E2Fs. Such conservation indicates a functional similarity between plant and animal E2Fs. E2F transcription factors that regulate meristem development act through common cis-elements, and regulate related (PCNA) genes (Kosugi and Ohashi, (2002) *Plant J.* 29: 45-59).

(3) The ABI5 gene (ABA insensitive 5) encodes a basic leucine zipper factor required for ABA response in the seed and vegetative tissues. Co-transformation experiments with ABI5 cDNA constructs in rice protoplasts resulted in specific transactivation of the ABA-inducible wheat, *Arabidopsis*, bean, and barley promoters. These results demonstrate that sequentially similar ABI5 transcription factors are key targets of a conserved ABA signaling pathway in diverse plants. (Gampala et al. (2001) *J. Biol. Chem.* 277: 1689-1694).

(4) Sequences of three *Arabidopsis* GAMYB-like genes were obtained on the basis of sequence similarity to GAMYB genes from barley, rice, and *L. temulentum*. These three *Arabadopsis* genes were determined to encode transcription factors (AtMYB33, AtMYB65, and AtMYB101) and could substitute for a barley GAMYB and control alpha-amylase expression (Gocal et al. (2001) *Plant Physiol.* 127: 1682-1693).

(5) The floral control gene LEAFY from *Arabidopsis* can dramatically accelerate flowering in numerous dictoyledonous plants. Constitutive expression of *Arabidopsis* LEAFY also caused early flowering in transgenic rice (a monocot), with a heading date that was 26-34 days earlier than that of wild-type plants. These observations indicate that floral regulatory genes from *Arabidopsis* are useful tools for heading date improvement in cereal crops (He et al. (2000) *Transgenic Res.* 9: 223-227).

(6) Bioactive gibberellins (GAs) are essential endogenous regulators of plant growth. GA signaling tends to be conserved across the plant kingdom. GA signaling is mediated via GAI, a nuclear member of the GRAS family of plant transcription factors. *Arabidopsis* GAI has been shown to function in rice to inhibit gibberellin response pathways (Fu et al. (2001) *Plant Cell* 13: 1791-1802).

(7) The *Arabidopsis* gene SUPERMAN (SUP), encodes a putative transcription factor that maintains the boundary between stamens and carpels. By over-expressing *Arabidopsis* SUP in rice, the effect of the gene's presence on whorl boundaries was shown to be conserved. This demonstrated that SUP is a conserved regulator of floral whorl boundaries and affects cell proliferation (Nandi et al. (2000) *Curr. Biol.* 10: 215-218).

(8) Maize, petunia and *Arabidopsis* myb transcription factors that regulate flavonoid biosynthesis are very genetically similar and affect the same trait in their native species, therefore sequence and function of these myb transcription factors correlate with each other in these diverse species (Borevitz et al. (2000) *Plant Cell* 12: 2383-2394).

(9) Wheat reduced height-1 (Rht-B1/Rht-D1) and maize dwarf-8 (d8) genes are orthologs of the *Arabidopsis* gibberellin insensitive (GAI) gene. Both of these genes have been used to produce dwarf grain varieties that have improved grain yield. These genes encode proteins that resemble nuclear transcription factors and contain an SH2-like domain, indicating that phosphotyrosine may

participate in gibberellin signaling. Transgenic rice plants containing a mutant GAI allele from *Arabidopsis* have been shown to produce reduced responses to gibberellin and are dwarfed, indicating that mutant GAI orthologs could be used to increase yield in a wide range of crop species (Peng et al. (1999) *Nature* 400: 256-261).

Transcription factors that are homologous to the listed sequences will typically share at least about 70% amino acid sequence identity in B domain. More closely related transcription factors can share at least about 75% or about 80% or about 90% or about 95% or about 98% or more sequence identity with the listed sequences, or with the listed sequences but excluding or outside a known consensus sequence or consensus DNA-binding site, or with the listed sequences excluding one or all conserved domains. Factors that are most closely related to the listed sequences share, e.g., at least about 85%, about 90% or about 95% or more % sequence identity to the listed sequences, or to the listed sequences but excluding or outside a known consensus sequence or consensus DNA-binding site or outside one or all conserved domain. At the nucleotide level, the sequences will typically share at least about 40% nucleotide sequence identity, preferably at least about 50%, about 60%, about 70% or about 80% sequence identity, and more preferably about 85%, about 90%, about 95% or about 97% or more sequence identity to one or more of the listed sequences, or to a listed sequence but excluding or outside a known consensus sequence or consensus DNA-binding site, or outside one or all conserved domain. The degeneracy of the genetic code enables major variations in the nucleotide sequence of a polynucleotide while maintaining the amino acid sequence of the encoded protein. B domains within the CCAAT-binding transcription factor family may exhibit a higher degree of sequence homology, such as at least 70% amino acid sequence identity including conservative substitutions, and preferably at least 80% sequence identity, and more preferably at least 85%, or at least about 86%, or at least about 87%, or at least about 88%, or at least about 90%, or at least about 95%, or at least about 98% sequence identity. Transcription factors that are homologous to the listed sequences should share at least 30%, or at least about 60%, or at least about 75%, or at least about 80%, or at least about 90%, or at least about 95% amino acid sequence identity over the entire length of the polypeptide or the homolog.

Percent identity can be determined electronically, e.g., by using the MEGALIGN program (DNASTAR, Inc. Madison, Wis.). The MEGALIGN program can create alignments between two or more sequences according to different methods, for example, the clustal method. (See, for example, Higgins and Sharp (1988) *Gene* 73: 237-244.) The clustal algorithm groups sequences into clusters by examining the distances between all pairs. The clusters are aligned pairwise and then in groups. Other alignment algorithms or programs may be used, including FASTA, BLAST, or ENTREZ, FASTA and BLAST, and which may be used to calculate percent similarity. These are available as a part of the

GCG sequence analysis package (University of Wisconsin, Madison, Wis.), and can be used with or without default settings. ENTREZ is available through the National Center for Biotechnology Information. In one embodiment, the percent identity of two sequences can be determined by the GCG program with a gap weight of 1, e.g., each amino acid gap is weighted as if it were a single amino acid or nucleotide mismatch between the two sequences (see USPN 6,262,333).

Other techniques for alignment are described in Methods in Enzymology, vol. 266, Computer Methods for Macromolecular Sequence Analysis (1996), ed. Doolittle, Academic Press, Inc., San Diego, Calif., USA. Preferably, an alignment program that permits gaps in the sequence is utilized to align the sequences. The Smith-Waterman is one type of algorithm that permits gaps in sequence alignments (see Shpaer (1997) *Methods Mol. Biol.* 70: 173-187). Also, the GAP program using the Needleman and Wunsch alignment method can be utilized to align sequences. An alternative search strategy uses MPSRCH software, which runs on a MASPAR computer. MPSRCH uses a Smith-Waterman algorithm to score sequences on a massively parallel computer. This approach improves ability to pick up distantly related matches, and is especially tolerant of small gaps and nucleotide sequence errors. Nucleic acid-encoded amino acid sequences can be used to search both protein and DNA databases.

The percentage similarity between two polypeptide sequences, e.g., sequence A and sequence B, is calculated by dividing the length of sequence A, minus the number of gap residues in sequence A, minus the number of gap residues in sequence B, into the sum of the residue matches between sequence A and sequence B, times one hundred. Gaps of low or of no similarity between the two amino acid sequences are not included in determining percentage similarity. Percent identity between polynucleotide sequences can also be counted or calculated by other methods known in the art, e.g., the Jotun Hein method. (See, e.g., Hein (1990) *Methods Enzymol.* 183: 626-645.) Identity between sequences can also be determined by other methods known in the art, e.g., by varying hybridization conditions (see US Patent Application No. 20010010913).

Thus, the invention provides methods for identifying a sequence similar or paralogous or orthologous or homologous to one or more polynucleotides as noted herein, or one or more target polypeptides encoded by the polynucleotides, or otherwise noted herein and may include linking or associating a given plant phenotype or gene function with a sequence. In the methods, a sequence database is provided (locally or across an internet or intranet) and a query is made against the sequence database using the relevant sequences herein and associated plant phenotypes or gene functions.

In addition, one or more polynucleotide sequences or one or more polypeptides encoded by the polynucleotide sequences may be used to search against a BLOCKS (Bairoch et al. (1997) *Nucleic Acids Res.* 25: 217-221), PFAM, and other databases which contain previously identified and

annotated motifs, sequences and gene functions. Methods that search for primary sequence patterns with secondary structure gap penalties (Smith et al. (1992) *Protein Engineering* 5: 35-51) as well as algorithms such as Basic Local Alignment Search Tool (BLAST; Altschul (1993) *J. Mol. Evol.* 36: 290-300; Altschul et al. (1990) *supra*), BLOCKS (Henikoff and Henikoff (1991) *Nucleic Acids Res.* 19: 6565-6572), Hidden Markov Models (HMM; Eddy (1996) *Curr. Opin. Str. Biol.* 6: 361-365; Sonnhammer et al. (1997) *Proteins* 28: 405-420), and the like, can be used to manipulate and analyze polynucleotide and polypeptide sequences encoded by polynucleotides. These databases, algorithms and other methods are well known in the art and are described in Ausubel et al. (1997; Short Protocols in Molecular Biology, John Wiley & Sons, New York, NY, unit 7.7) and in Meyers (1995; Molecular Biology and Biotechnology, Wiley VCH, New York, NY, p 856-853).

A further method for identifying or confirming that specific homologous sequences control the same function is by comparison of the transcript profile(s) obtained upon overexpression or knockout of two or more related transcription factors. Since transcript profiles are diagnostic for specific cellular states, one skilled in the art will appreciate that genes that have a highly similar transcript profile (e.g., with greater than 50% regulated transcripts in common, more preferably with greater than 70% regulated transcripts in common, most preferably with greater than 90% regulated transcripts in common) will have highly similar functions. Fowler et al. (2002, Plant Cell, 14: 1675-79) have shown that three paralogous AP2 family genes (CBF1, CBF2 and CBF3), each of which is induced upon cold treatment, and each of which can condition improved freezing tolerance, have highly similar transcript profiles. Once a transcription factor has been shown to provide a specific function, its transcript profile becomes a diagnostic tool to determine whether putative paralogs or orthologs have the same function.

Furthermore, methods using manual alignment of sequences similar or homologous to one or more polynucleotide sequences or one or more polypeptides encoded by the polynucleotide sequences may be used to identify regions of similarity and B domains. Such manual methods are well-known of those of skill in the art and can include, for example, comparisons of tertiary structure between a polypeptide sequence encoded by a polynucleotide which comprises a known function with a polypeptide sequence encoded by a polynucleotide sequence which has a function not yet determined. Such examples of tertiary structure may comprise predicted alpha helices, beta-sheets, amphipathic helices, leucine zipper motifs, zinc finger motifs, proline-rich regions, cysteine repeat motifs, and the like.

Orthologs and paralogs of presently disclosed transcription factors may be cloned using compositions provided by the present invention according to methods well known in the art. cDNAs can be cloned using mRNA from a plant cell or tissue that expresses one of the present transcription factors. Appropriate mRNA sources may be identified by interrogating Northern blots with probes

designed from the present transcription factor sequences, after which a library is prepared from the mRNA obtained from a positive cell or tissue. Transcription factor-encoding cDNA is then isolated using, for example, PCR, using primers designed from a presently disclosed transcription factor gene sequence, or by probing with a partial or complete cDNA or with one or more sets of degenerate probes based on the disclosed sequences. The cDNA library may be used to transform plant cells. Expression of the cDNAs of interest is detected using, for example, methods disclosed herein such as microarrays, Northern blots, quantitative PCR, or any other technique for monitoring changes in expression. Genomic clones may be isolated using similar techniques to those.

In addition to the Sequences listed in the Sequence Listing, the invention encompasses isolated nucleotide sequences that are sequentially and structurally similar to G481, G482, and G485, SEQ ID NO: 1, 3, and 5, and function in a plant in a manner similar to G481, G482 and G485 by regulating abiotic stress tolerance. The nucleotide sequences of G481 and G485 are 88% and 82% identical to the polynucleotide sequence of G482, respectively. Since all three polynucleotide sequences are phylogenetically related, sequentially similar, and have been shown to regulate abiotic stress tolerance, one skilled in the art would predict that other similar, phylogenetically related sequences would also regulate abiotic stress tolerance. A sequence that was 99.5% identical (861 of 865 bases) to G482 has been taught by Edwards et al., ((1998) *Plant Physiol.* 117: 1015-1022), but with no analysis of the function of this gene.

The present invention is also directed to polypeptide encoded by isolated nucleic acid that are similar tc G481, G482 and G485, vectors comprising isolated nucleic acid that are similar to G481, G482 and G485, and transgenic plants transformed with these isolated nucleic acids.

Identifying Polynucleotides or Nucleic Acids by Hybridization

Polynucleotides homologous to the sequences illustrated in the Sequence Listing and tables can be identified, e.g., by hybridization to each other under stringent or under highly stringent conditions. Single stranded polynucleotides hybridize when they associate based on a variety of well characterized physical-chemical forces, such as hydrogen bonding, solvent exclusion, base stacking and the like. The stringency of a hybridization reflects the degree of sequence identity of the nucleic acids involved, such that the higher the stringency, the more similar are the two polynucleotide strands. Stringency is influenced by a variety of factors, including temperature, salt concentration and composition, organic and non-organic additives, solvents, etc. present in both the hybridization and wash solutions and incubations (and number thereof), as described in more detail in the references cited above.

Encompassed by the invention are polynucleotide sequences that are capable of hybridizing to

the claimed polynucleotide sequences, including any of the transcription factor polynucleotides within the Sequence Listing, and fragments thereof under various conditions of stringency (See, for example, Wahl and Berger (1987) *Methods Enzymol.* 152: 399-407; and Kimmel (1987) *Methods Enzymol.* 152: 507-511). In addition to the nucleotide sequences in the Sequence Listing, full-length cDNA, orthologs, and paralogs of the present nucleotide sequences may be identified and isolated using well-known methods. The cDNA libraries, orthologs, and paralogs of the present nucleotide sequences may be screened using hybridization methods to determine their utility as hybridization target or amplification probes.

With regard to hybridization, conditions that are highly stringent, and means for achieving them, are well known in the art. See, for example, Sambrook et al. (1989) "*Molecular Cloning: A Laboratory Manual*" (2nd ed., Cold Spring Harbor Laboratory); Berger and Kimmel, eds., (1987) "Guide to Molecular Cloning Techniques", In *Methods in Enzymology*:152: 467-469; and Anderson and Young (1985) "Quantitative Filter Hybridisation." In: Hames and Higgins, ed., Nucleic Acid Hybridisation, A Practical Approach. Oxford, IRL Press, 73–111.

Stability of DNA duplexes is affected by such factors as base composition, length, and degree of base pair mismatch. Hybridization conditions may be adjusted to allow DNAs of different sequence relatedness to hybridize. The melting temperature (T_m) is defined as the temperature when 50% of the duplex molecules have dissociated into their constituent single strands. The melting temperature of a perfectly matched duplex, where the hybridization buffer contains formamide as a denaturing agent, may be estimated by the following equations:

(I) DNA-DNA:

$T_m(^\circ C)=81.5+16.6(\log [Na+])+0.41(\%\ G+C)- 0.62(\%\ \text{formamide})-500/L$

(II) DNA-RNA:

$T_m(^\circ C)=79.8+18.5(\log [Na+])+0.58(\%\ G+C)+ 0.12(\%G+C)^2 - 0.5(\%\ \text{formamide}) - 820/L$

(III) RNA-RNA:

$T_m(^\circ C)=79.8+18.5(\log [Na+])+0.58(\%\ G+C)+ 0.12(\%G+C)^2 - 0.35(\%\ \text{formamide}) - 820/L$

where L is the length of the duplex formed, [Na+] is the molar concentration of the sodium ion in the hybridization or washing solution, and % G+C is the percentage of (guanine+cytosine) bases in the hybrid. For imperfectly matched hybrids, approximately 1° C is required to reduce the melting temperature for each 1% mismatch.

Hybridization experiments are generally conducted in a buffer of pH between 6.8 to 7.4, although the rate of hybridization is nearly independent of pH at ionic strengths likely to be used in the hybridization buffer (Anderson et al. (1985) *supra*). In addition, one or more of the following may be used to reduce non-specific hybridization: sonicated salmon sperm DNA or another non-complementary DNA, bovine serum albumin, sodium pyrophosphate, sodium dodecylsulfate (SDS), polyvinyl-pyrrolidone, ficoll and Denhardt's solution. Dextran sulfate and polyethylene glycol 6000 act to exclude DNA from solution, thus raising the effective probe DNA concentration and the hybridization signal within a given unit of time. In some instances, conditions of even greater stringency may be desirable or required to reduce non-specific and/or background hybridization. These conditions may be created with the use of higher temperature, lower ionic strength and higher concentration of a denaturing agent such as formamide.

Stringency conditions can be adjusted to screen for moderately similar fragments such as homologous sequences from distantly related organisms, or to highly similar fragments such as genes that duplicate functional enzymes from closely related organisms. The stringency can be adjusted either during the hybridization step or in the post-hybridization washes. Salt concentration, formamide concentration, hybridization temperature and probe lengths are variables that can be used to alter stringency (as described by the formula above). As a general guidelines high stringency is typically performed at T_m-5° C to T_m-20° C, moderate stringency at T_m-20° C to T_m-35° C and low stringency at T_m-35° C to T_m-50° C for duplex >150 base pairs. Hybridization may be performed at low to moderate stringency (25-50° C below T_m), followed by post-hybridization washes at increasing stringencies. Maximum rates of hybridization in solution are determined empirically to occur at T_m-25° C for DNA-DNA duplex and T_m-15° C for RNA-DNA duplex. Optionally, the degree of dissociation may be assessed after each wash step to determine the need for subsequent, higher stringency wash steps.

High stringency conditions may be used to select for nucleic acid sequences with high degrees of identity to the disclosed sequences. An example of stringent hybridization conditions obtained in a filter-based method such as a Southern or northern blot for hybridization of complementary nucleic acids that have more than 100 complementary residues is about 5°C to 20°C lower than the thermal melting point (T_m) for the specific sequence at a defined ionic strength and pH. Conditions used for hybridization may include about 0.02 M to about 0.15 M sodium chloride, about 0.5% to about 5% casein, about 0.02% SDS or about 0.1% N-laurylsarcosine, about 0.001 M to about 0.03 M sodium citrate, at hybridization temperatures between about 50° C and about 70° C. More preferably, high stringency conditions are about 0.02 M sodium chloride, about 0.5% casein, about 0.02% SDS, about 0.001 M sodium citrate, at a temperature of about 50° C. Nucleic acid molecules that hybridize under stringent conditions will typically hybridize to a probe based on either the entire DNA molecule or

selected portions, e.g., to a unique subsequence, of the DNA.

Stringent salt concentration will ordinarily be less than about 750 mM NaCl and 75 mM trisodium citrate. Increasingly stringent conditions may be obtained with less than about 500 mM NaCl and 50 mM trisodium citrate, to even greater stringency with less than about 250 mM NaCl and 25 mM trisodium citrate. Low stringency hybridization can be obtained in the absence of organic solvent, e.g., formamide, whereas high stringency hybridization may be obtained in the presence of at least about 35% formamide, and more preferably at least about 50% formamide. Stringent temperature conditions will ordinarily include temperatures of at least about 30° C, more preferably of at least about 37° C, and most preferably of at least about 42° C with formamide present. Varying additional parameters, such as hybridization time, the concentration of detergent, e.g., sodium dodecyl sulfate (SDS) and ionic strength, are well known to those skilled in the art. Various levels of stringency are accomplished by combining these various conditions as needed.

The washing steps that follow hybridization may also vary in stringency; the post-hybridization wash steps primarily determine hybridization specificity, with the most critical factors being temperature and the ionic strength of the final wash solution. Wash stringency can be increased by decreasing salt concentration or by increasing temperature. Stringent salt concentration for the wash steps will preferably be less than about 30 mM NaCl and 3 mM trisodium citrate, and most preferably less than about 15 mM NaCl and 1.5 mM trisodium citrate.

Thus, hybridization and wash conditions that may be used to bind and remove polynucleotides with less than the desired homology to the nucleic acid sequences or their complements that encode the present transcription factors include, for example:

6X SSC at 65° C;

50% formamide, 4X SSC at 42° C; or

0.5X SSC, 0.1% SDS at 65° C;

with, for example, two wash steps of 10 - 30 minutes each. . Useful variations on these conditions will be readily apparent to those skilled in the art.

A person of skill in the art would not expect substantial variation among polynucleotide species encompassed within the scope of the present invention because the highly stringent conditions set forth in the above formulae yield structurally similar polynucleotides.

If desired, one may employ wash steps of even greater stringency, including about 0.2x SSC, 0.1% SDS at 65° C and washing twice, each wash step being about 30 min, or about 0.1 x SSC, 0.1% SDS at 65° C and washing twice for 30 min. The temperature for the wash solutions will ordinarily be at least about 25° C, and for greater stringency at least about 42° C. Hybridization stringency may be increased further by using the same conditions as in the hybridization steps, with the wash temperature

raised about 3° C to about 5° C, and stringency may be increased even further by using the same conditions except the wash temperature is raised about 6° C to about 9° C. For identification of less closely related homologs, wash steps may be performed at a lower temperature, e.g., 50° C.

An example of a low stringency wash step employs a solution and conditions of at least 25° C in 30 mM NaCl, 3 mM trisodium citrate, and 0.1% SDS over 30 min. Greater stringency may be obtained at 42° C in 15 mM NaCl, with 1.5 mM trisodium citrate, and 0.1% SDS over 30 min. Even higher stringency wash conditions are obtained at 65° C -68° C in a solution of 15 mM NaCl, 1.5 mM trisodium citrate, and 0.1% SDS. Wash procedures will generally employ at least two final wash steps. Additional variations on these conditions will be readily apparent to those skilled in the art (see, for example, US Patent Application No. 20010010913).

Stringency conditions can be selected such that an oligonucleotide that is perfectly complementary to the coding oligonucleotide hybridizes to the coding oligonucleotide with at least about a 5-10x higher signal to noise ratio than the ratio for hybridization of the perfectly complementary oligonucleotide to a nucleic acid encoding a transcription factor known as of the filing date of the application. It may be desirable to select conditions for a particular assay such that a higher signal to noise ratio, that is, about 15x or more, is obtained. Accordingly, a subject nucleic acid will hybridize to a unique coding oligonucleotide with at least a 2x or greater signal to noise ratio as compared to hybridization of the coding oligonucleotide to a nucleic acid encoding known polypeptide. The particular signal will depend on the label used in the relevant assay, e.g., a fluorescent label, a colorimetric label, a radioactive label, or the like. Labeled hybridization or PCR probes for detecting related polynucleotide sequences may be produced by oligolabeling, nick translation, end-labeling, or PCR amplification using a labeled nucleotide.

Encompassed by the invention are polynucleotide sequences that are capable of hybridizing to the present polynucleotide sequences, and, in particular, to SEQ ID NOs: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, polynucleotides that encode polypeptide SEQ ID NOs: 29-32, and fragments thereof under various conditions of stringency. (See, e.g., Wahl and Berger (1987) *Methods Enzymol.* 152: 399-407; Kimmel (1987) *Methods Enzymol.* 152: 507-511.) Estimates of homology are provided by either DNA-DNA or DNA-RNA hybridization under conditions of stringency as is well understood by those skilled in the art (Hames and Higgins, Eds. (1985) Nucleic Acid Hybridisation, IRL Press, Oxford, U.K.). Stringency conditions can be adjusted to screen for moderately similar fragments, such as homologous sequences from distantly related organisms, to highly similar fragments, such as genes that duplicate functional enzymes from closely related organisms. Post-hybridization washes determine stringency conditions.

Identifying Polynucleotides or Nucleic Acids with Expression Libraries

In addition to hybridization methods, transcription factor homolog polypeptides can be obtained by screening an expression library using antibodies specific for one or more transcription factors. With the provision herein of the disclosed transcription factor, and transcription factor homolog nucleic acid sequences, the encoded polypeptide(s) can be expressed and purified in a heterologous expression system (e.g., *E. coli*) and used to raise antibodies (monoclonal or polyclonal) specific for the polypeptide(s) in question. Antibodies can also be raised against synthetic peptides derived from transcription factor, or transcription factor homolog, amino acid sequences. Methods of raising antibodies are well known in the art and are described in Harlow and Lane (1988), Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory, New York. Such antibodies can then be used to screen an expression library produced from the plant from which it is desired to clone additional transcription factor homologs, using the methods described above. The selected cDNAs can be confirmed by sequencing and enzymatic activity.

Sequence Variations

It will readily be appreciated by those of skill in the art, that any of a variety of polynucleotide sequences are capable of encoding the transcription factors and transcription factor homolog polypeptides of the invention. Due to the degeneracy of the genetic code, many different polynucleotides can encode identical and/or substantially similar polypeptides in addition to those sequences illustrated in the Sequence Listing. Nucleic acids having a sequence that differs from the sequences shown in the Sequence Listing, or complementary sequences, that encode functionally equivalent peptides (i.e., peptides having some degree of equivalent or similar biological activity) but differ in sequence from the sequence shown in the Sequence Listing due to degeneracy in the genetic code, are also within the scope of the invention.

Altered polynucleotide sequences encoding polypeptides include those sequences with deletions, insertions, or substitutions of different nucleotides, resulting in a polynucleotide encoding a polypeptide with at least one functional characteristic of the instant polypeptides. Included within this definition are polymorphisms which may or may not be readily detectable using a particular oligonucleotide probe of the polynucleotide encoding the instant polypeptides, and improper or unexpected hybridization to allelic variants, with a locus other than the normal chromosomal locus for the polynucleotide sequence encoding the instant polypeptides.

Allelic variant refers to any of two or more alternative forms of a gene occupying the same chromosomal locus. Allelic variation arises naturally through mutation, and may result in phenotypic

polymorphism within populations. Gene mutations can be silent (i.e., no change in the encoded polypeptide) or may encode polypeptides having altered amino acid sequence. The term allelic variant is also used herein to denote a protein encoded by an allelic variant of a gene. Splice variant refers to alternative forms of RNA transcribed from a gene. Splice variation arises naturally through use of alternative splicing sites within a transcribed RNA molecule, or less commonly between separately transcribed RNA molecules, and may result in several mRNAs transcribed from the same gene. Splice variants may encode polypeptides having altered amino acid sequence. The term splice variant is also used herein to denote a protein encoded by a splice variant of an mRNA transcribed from a gene.

Those skilled in the art would recognize that, for example, G482, SEQ ID NO: 4, represents a single transcription factor; allelic variation and alternative splicing may be expected to occur. Allelic variants of SEQ ID NO: 3 can be cloned by probing cDNA or genomic libraries from different individual organisms according to standard procedures. Allelic variants of the DNA sequence shown in SEQ ID NO: 3, including those containing silent mutations and those in which mutations result in amino acid sequence changes, are within the scope of the present invention, as are proteins which are allelic variants of SEQ ID NO: 4. cDNAs generated from alternatively spliced mRNAs, which retain the properties of the transcription factor are included within the scope of the present invention, as are polypeptides encoded by such cDNAs and mRNAs. Allelic variants and splice variants of these sequences can be cloned by probing cDNA or genomic libraries from different individual organisms or tissues according to standard procedures known in the art (see USPN 6,388,064).

Thus, in addition to the sequences set forth in the Sequence Listing, the invention also encompasses related nucleic acid molecules that include allelic or splice variants SEQ ID NO: 1, 3, 5, 7, 9, 11-21, 27-52, 55, 57, 59, 61, 63, 65, 67, 69, 71, 75, 77, and 79, and include sequences which are complementary to these nucleotide sequences. Related nucleic acid molecules also include nucleotide sequences encoding a polypeptide comprising or consisting essentially of a substitution, modification, addition and/or deletion of one or more amino acid residues compared to the polypeptide as set forth in any of SEQ ID NOs: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 29, 30, 31, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92 and 94. Such related polypeptides may comprise, for example, additions and/or deletions of one or more N-linked or O-linked glycosylation sites, or an addition and/or a deletion of one or more cysteine residues.

For example, Table 2 illustrates, e.g., that the codons AGC, AGT, TCA, TCC, TCG, and TCT all encode the same amino acid: serine. Accordingly, at each position in the sequence where there is a codon encoding serine, any of the above trinucleotide sequences can be used without altering the encoded polypeptide.

Table 2

Amino acid			Possible Codons					
Alanine	Ala	A	GCA	GCC	GCG	GCT		
Cysteine	Cys	C	TGC	TGT				
Aspartic acid	Asp	D	GAC	GAT				
Glutamic acid	Glu	E	GAA	GAG				
Phenylalanine	Phe	F	TTC	TTT				
Glycine	Gly	G	GGA	GGC	GGG	GGT		
Histidine	His	H	CAC	CAT				
Isoleucine	Ile	I	ATA	ATC	ATT			
Lysine	Lys	K	AAA	AAG				
Leucine	Leu	L	TTA	TTG	CTA	CTC	CTG	CTT
Methionine	Met	M	ATG					
Asparagine	Asn	N	AAC	AAT				
Proline	Pro	P	CCA	CCC	CCG	CCT		
Glutamine	Gln	Q	CAA	CAG				
Arginine	Arg	R	AGA	AGG	CGA	CGC	CGG	CGT
Serine	Ser	S	AGC	AGT	TCA	TCC	TCG	TCT
Threonine	Thr	T	ACA	ACC	ACG	ACT		
Valine	Val	V	GTA	GTC	GTG	GTT		
Tryptophan	Trp	W	TGG					
Tyrosine	Tyr	Y	TAC	TAT				

Sequence alterations that do not change the amino acid sequence encoded by the polynucleotide are termed "silent" variations. With the exception of the codons ATG and TGG, encoding methionine and tryptophan, respectively, any of the possible codons for the same amino acid can be substituted by a variety of techniques, e.g., site-directed mutagenesis, available in the art. Accordingly, any and all such variations of a sequence selected from the above table are a feature of the invention.

In addition to silent variations, other conservative variations that alter one, or a few amino acids in the encoded polypeptide, can be made without altering the function of the polypeptide, these conservative variants are, likewise, a feature of the invention.

For example, substitutions, deletions and insertions introduced into the sequences provided in the Sequence Listing, are also envisioned by the invention. Such sequence modifications can be engineered into a sequence by site-directed mutagenesis (Wu (ed.) *Methods Enzymol.* (1993) vol. 217, Academic Press) or the other methods noted below. Amino acid substitutions are typically of single residues; insertions usually will be on the order of about from 1 to 10 amino acid residues; and deletions will range about from 1 to 30 residues. In preferred embodiments, deletions or insertions are made in adjacent pairs, e.g., a deletion of two residues or insertion of two residues. Substitutions, deletions, insertions or any combination thereof can be combined to arrive at a sequence. The

mutations that are made in the polynucleotide encoding the transcription factor should not place the sequence out of reading frame and should not create complementary regions that could produce secondary mRNA structure. Preferably, the polypeptide encoded by the DNA performs the desired function.

Conservative substitutions are those in which at least one residue in the amino acid sequence has been removed and a different residue inserted in its place. Such substitutions generally are made in accordance with the Table 3 when it is desired to maintain the activity of the protein. Table 3 shows amino acids which can be substituted for an amino acid in a protein and which are typically regarded as conservative substitutions.

Table 3

Residue	Conservative Substitutions
Ala	Ser
Arg	Lys
Asn	Gln; His
Asp	Glu
Gln	Asn
Cys	Ser
Glu	Asp
Gly	Pro
His	Asn; Gln
Ile	Leu, Val
Leu	Ile; Val
Lys	Arg; Gln
Met	Leu; Ile
Phe	Met; Leu; Tyr
Ser	Thr; Gly
Thr	Ser; Val
Trp	Tyr
Tyr	Trp; Phe
Val	Ile; Leu

Similar substitutions are those in which at least one residue in the amino acid sequence has been removed and a different residue inserted in its place. Such substitutions generally are made in accordance with the Table 4 when it is desired to maintain the activity of the protein. Table 4 shows amino acids which can be substituted for an amino acid in a protein and which are typically regarded as structural and functional substitutions. For example, a residue in column 1 of Table 4 may be substituted with a residue in column 2; in addition, a residue in column 2 of Table 4 may be substituted with the residue of column 1.

Table 4

Residue	Similar Substitutions
Ala	Ser; Thr; Gly; Val; Leu; Ile
Arg	Lys; His; Gly
Asn	Gln; His; Gly; Ser; Thr
Asp	Glu, Ser; Thr
Gln	Asn; Ala
Cys	Ser; Gly
Glu	Asp
Gly	Pro; Arg
His	Asn; Gln; Tyr; Phe; Lys; Arg
Ile	Ala; Leu; Val; Gly; Met
Leu	Ala; Ile; Val; Gly; Met
Lys	Arg; His; Gln; Gly; Pro
Met	Leu; Ile; Phe
Phe	Met; Leu; Tyr; Trp; His; Val; Ala
Ser	Thr; Gly; Asp; Ala; Val; Ile; His
Thr	Ser; Val; Ala; Gly
Trp	Tyr; Phe; His
Tyr	Trp; Phe; His
Val	Ala; Ile; Leu; Gly; Thr; Ser; Glu

Substitutions that are less conservative than those in Table 3 can be selected by picking residues that differ more significantly in their effect on maintaining (a) the structure of the polypeptide backbone in the area of the substitution, for example, as a sheet or helical conformation, (b) the charge

or hydrophobicity of the molecule at the target site, or (c) the bulk of the side chain. The substitutions which in general are expected to produce the greatest changes in protein properties will be those in which (a) a hydrophilic residue, e.g., seryl or threonyl, is substituted for (or by) a hydrophobic residue, e.g., leucyl, isoleucyl, phenylalanyl, valyl or alanyl; (b) a cysteine or proline is substituted for (or by) any other residue; (c) a residue having an electropositive side chain, e.g., lysyl, arginyl, or histidyl, is substituted for (or by) an electronegative residue, e.g., glutamyl or aspartyl; or (d) a residue having a bulky side chain, e.g., phenylalanine, is substituted for (or by) one not having a side chain, e.g., glycine.

Further Modifying Sequences of the Invention – Mutation/Forced Evolution

In addition to generating silent or conservative substitutions as noted, above, the present invention optionally includes methods of modifying the sequences of the Sequence Listing. In the methods, nucleic acid or protein modification methods are used to alter the given sequences to produce new sequences and/or to chemically or enzymatically modify given sequences to change the properties of the nucleic acids or proteins.

Thus, in one embodiment, given nucleic acid sequences are modified, e.g., according to standard mutagenesis or artificial evolution methods to produce modified sequences. The modified sequences may be created using purified natural polynucleotides isolated from any organism or may be synthesized from purified compositions and chemicals using chemical means well know to those of skill in the art. For example, Ausubel, *supra*, provides additional details on mutagenesis methods. Artificial forced evolution methods are described, for example, by Stemmer (1994) *Nature* 370: 389-391, Stemmer (1994) *Proc. Natl. Acad. Sci.* 91: 10747-10751, and US Patents 5,811,238, 5,837,500, and 6,242,568. Methods for engineering synthetic transcription factors and other polypeptides are described, for example, by Zhang et al. (2000) *J. Biol. Chem.* 275: 33850-33860, Liu et al. (2001) *J. Biol. Chem.* 276: 11323-11334, and Isalan et al. (2001) *Nature Biotechnol.* 19: 656-660. Many other mutation and evolution methods are also available and expected to be within the skill of the practitioner.

Similarly, chemical or enzymatic alteration of expressed nucleic acids and polypeptides can be performed by standard methods. For example, sequence can be modified by addition of lipids, sugars, peptides, organic or inorganic compounds, by the inclusion of modified nucleotides or amino acids, or the like. For example, protein modification techniques are illustrated in Ausubel, *supra*. Further details on chemical and enzymatic modifications can be found herein. These modification methods can be used to modify any given sequence, or to modify any sequence produced by the various mutation and artificial evolution modification methods noted herein.

Accordingly, the invention provides for modification of any given nucleic acid by mutation, evolution, chemical or enzymatic modification, or other available methods, as well as for the products produced by practicing such methods, e.g., using the sequences herein as a starting substrate for the various modification approaches.

For example, optimized coding sequence containing codons preferred by a particular prokaryotic or eukaryotic host can be used e.g., to increase the rate of translation or to produce recombinant RNA transcripts having desirable properties, such as a longer half-life, as compared with transcripts produced using a non-optimized sequence. Translation stop codons can also be modified to reflect host preference. For example, preferred stop codons for *Saccharomyces cerevisiae* and mammals are TAA and TGA, respectively. The preferred stop codon for monocotyledonous plants is TGA, whereas insects and *E. coli* prefer to use TAA as the stop codon.

The polynucleotide sequences of the present invention can also be engineered in order to alter a coding sequence for a variety of reasons, including but not limited to, alterations which modify the sequence to facilitate cloning, processing and/or expression of the gene product. For example, alterations are optionally introduced using techniques which are well known in the art, e.g., site-directed mutagenesis, to insert new restriction sites, to alter glycosylation patterns, to change codon preference, to introduce splice sites, etc.

Furthermore, a fragment or domain derived from any of the polypeptides of the invention can be combined with domains derived from other transcription factors or synthetic domains to modify the biological activity of a transcription factor. For instance, a DNA-binding domain derived from a transcription factor of the invention can be combined with the activation domain of another transcription factor or with a synthetic activation domain. A transcription activation domain assists in initiating transcription from a DNA-binding site. Examples include the transcription activation region of VP16 or GAL4 (Moore et al. (1998) *Proc. Natl. Acad. Sci.* 95: 376-381; Aoyama et al. (1995) *Plant Cell* 7: 1773-1785), peptides derived from bacterial sequences (Ma and Ptashne (1987) *Cell* 51: 113-119) and synthetic peptides (Giniger and Ptashne (1987) *Nature* 330: 670-672).

Expression and Modification of Polypeptides

Typically, polynucleotide sequences of the invention are incorporated into recombinant DNA (or RNA) molecules that direct expression of polypeptides of the invention in appropriate host cells, transgenic plants, in vitro translation systems, or the like. Due to the inherent degeneracy of the genetic code, nucleic acid sequences which encode substantially the same or a functionally equivalent amino acid sequence can be substituted for any listed sequence to provide for cloning and expressing the relevant homolog.

The transgenic plants of the present invention comprising recombinant polynucleotide sequences are generally derived from parental plants, which may themselves be non-transformed (or non-transgenic) plants. These transgenic plants may either have a transcription factor gene "knocked out" (for example, with a genomic insertion by homologous recombination, an antisense or ribozyme construct) or expressed to a normal or wild-type extent. However, overexpressing transgenic "progeny" plants will exhibit greater mRNA levels, wherein the mRNA encodes a transcription factor, that is, a DNA-binding protein that is capable of binding to a DNA regulatory sequence and inducing transcription, and preferably, expression of a plant trait gene. Preferably, the mRNA expression level will be at least three-fold greater than that of the parental plant, or more preferably at least ten-fold greater mRNA levels compared to said parental plant, and most preferably at least fifty-fold greater compared to said parental plant.

Vectors, Promoters, and Expression Systems

The present invention includes recombinant constructs comprising one or more of the nucleic acid sequences herein. The constructs typically comprise a vector, such as a plasmid, a cosmid, a phage, a virus (e.g., a plant virus), a bacterial artificial chromosome (BAC), a yeast artificial chromosome (YAC), or the like, into which a nucleic acid sequence of the invention has been inserted, in a forward or reverse orientation. In a preferred aspect of this embodiment, the construct further comprises regulatory sequences, including, for example, a promoter, operably linked to the sequence. Large numbers of suitable vectors and promoters are known to those of skill in the art, and are commercially available.

General texts that describe molecular biological techniques useful herein, including the use and production of vectors, promoters and many other relevant topics, include Berger, Sambrook, *supra* and Ausubel, *supra*. Any of the identified sequences can be incorporated into a cassette or vector, e.g., for expression in plants. A number of expression vectors suitable for stable transformation of plant cells or for the establishment of transgenic plants have been described including those described in Weissbach and Weissbach (1989) Methods for Plant Molecular Biology, Academic Press, and Gelvin et al. (1990) Plant Molecular Biology Manual, Kluwer Academic Publishers. Specific examples include those derived from a Ti plasmid of *Agrobacterium tumefaciens*, as well as those disclosed by Herrera-Estrella et al. (1983) *Nature* 303: 209, Bevan (1984) *Nucleic Acids Res.* 12: 8711-8721, Klee (1985) *Bio/Technology* 3: 637-642, for dicotyledonous plants.

Alternatively, non-Ti vectors can be used to transfer the DNA into monocotyledonous plants and cells by using free DNA delivery techniques. Such methods can involve, for example, the use of liposomes, electroporation, microprojectile bombardment, silicon carbide whiskers, and viruses. By

using these methods transgenic plants such as wheat, rice (Christou (1991) *Bio/Technology* 9: 957-962) and corn (Gordon-Kamm (1990) *Plant Cell* 2: 603-618) can be produced. An immature embryo can also be a good target tissue for monocots for direct DNA delivery techniques by using the particle gun (Weeks et al. (1993) *Plant Physiol.* 102: 1077-1084; Vasil (1993) *Bio/Technology* 10: 667-674; Wan and Lemeaux (1994) *Plant Physiol.* 104: 37-48, and for *Agrobacterium*-mediated DNA transfer (Ishida et al. (1996) *Nature Biotechnol.* 14: 745-750).

Typically, plant transformation vectors include one or more cloned plant coding sequence (genomic or cDNA) under the transcriptional control of 5' and 3' regulatory sequences and a dominant selectable marker. Such plant transformation vectors typically also contain a promoter (e.g., a regulatory region controlling inducible or constitutive, environmentally-or developmentally-regulated, or cell- or tissue-specific expression), a transcription initiation start site, an RNA processing signal (such as intron splice sites), a transcription termination site, and/or a polyadenylation signal.

A potential utility for the transcription factor polynucleotides disclosed herein is the isolation of promoter elements from these genes that can be used to program expression in plants of any genes. Each transcription factor gene disclosed herein is expressed in a unique fashion, as determined by promoter elements located upstream of the start of translation, and additionally within an intron of the transcription factor gene or downstream of the termination codon of the gene. As is well known in the art, for a significant portion of genes, the promoter sequences are located entirely in the region directly upstream of the start of translation. In such cases, typically the promoter sequences are located within 2.0 kb of the start of translation, or within 1.5 kb of the start of translation, frequently within 1.0 kb of the start of translation, and sometimes within 0.5 kb of the start of translation.

The promoter sequences can be isolated according to methods known to one skilled in the art.

Examples of constitutive plant promoters which can be useful for expressing the TF sequence include: the cauliflower mosaic virus (CaMV) 35S promoter, which confers constitutive, high-level expression in most plant tissues (*see,* e.g., Odell et al. (1985) *Nature* 313: 810-812); the nopaline synthase promoter (An et al. (1988) *Plant Physiol.* 88: 547-552); and the octopine synthase promoter (Fromm et al. (1989) *Plant Cell* 1: 977-984).

The transcription factors of the invention may be operably linked with a specific promoter that causes the transcription factor to be expressed in response to environmental, tissue-specific or temporal signals. A variety of plant gene promoters that regulate gene expression in response to environmental, hormonal, chemical, developmental signals, and in a tissue-active manner can be used for expression of a TF sequence in plants. Choice of a promoter is based largely on the phenotype of interest and is determined by such factors as tissue (e.g., seed, fruit, root, pollen, vascular tissue, flower, carpel, etc.), inducibility (e.g., in response to wounding, heat, cold, drought, light, pathogens, etc.), timing,

developmental stage, and the like. Numerous known promoters have been characterized and can favorably be employed to promote expression of a polynucleotide of the invention in a transgenic plant or cell of interest. For example, tissue specific promoters include: seed-specific promoters (such as the napin, phaseolin or DC3 promoter described in US Pat. No. 5,773,697), fruit-specific promoters that are active during fruit ripening (such as the dru 1 promoter (US Pat. No. 5,783,393), or the 2A11 promoter (US Pat. No. 4,943,674) and the tomato polygalacturonase promoter (Bird et al. (1988) *Plant Mol. Biol.* 11: 651-662), root-specific promoters, such as those disclosed in US Patent Nos. 5,618,988, 5,837,848 and 5,905,186, pollen-active promoters such as PTA29, PTA26 and PTA13 (US Pat. No. 5,792,929), promoters active in vascular tissue (Ringli and Keller (1998) *Plant Mol. Biol.* 37: 977-988), flower-specific (Kaiser et al. (1995) *Plant Mol. Biol.* 28: 231-243), pollen (Baerson et al. (1994) *Plant Mol. Biol.* 26: 1947-1959), carpels (Ohl et al. (1990) *Plant Cell* 2: 837-848), pollen and ovules (Baerson et al. (1993) *Plant Mol. Biol.* 22: 255-267), auxin-inducible promoters (such as that described in van der Kop et al. (1999) *Plant Mol. Biol.* 39: 979-990 or Baumann et al., (1999) *Plant Cell* 11: 323-334), cytokinin-inducible promoter (Guevara-Garcia (1998) *Plant Mol. Biol.* 38: 743-753), promoters responsive to gibberellin (Shi et al. (1998) *Plant Mol. Biol.* 38: 1053-1060, Willmott et al. (1998) *Plant Molec. Biol.* 38: 817-825) and the like. Additional promoters are those that elicit expression in response to heat (Ainley et al. (1993) *Plant Mol. Biol.* 22: 13-23), light (e.g., the pea rbcS-3A promoter, Kuhlemeier et al. (1989) *Plant Cell* 1: 471-478, and the maize rbcS promoter, Schaffner and Sheen (1991) *Plant Cell* 3: 997-1012); wounding (e.g., *wunI*, Siebertz et al. (1989) *Plant Cell* 1: 961-968); pathogens (such as the PR-1 promoter described in Buchel et al. (1999) *Plant Mol. Biol.* 40: 387-396, and the PDF1.2 promoter described in Manners et al. (1998) *Plant Mol. Biol.* 38: 1071-1080), and chemicals such as methyl jasmonate or salicylic acid (Gatz (1997) *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 48: 89-108). In addition, the timing of the expression can be controlled by using promoters such as those acting at senescence (Gan and Amasino (1995) *Science* 270: 1986-1988); or late seed development (Odell et al. (1994) *Plant Physiol.* 106: 447-458).

Plant expression vectors can also include RNA processing signals that can be positioned within, upstream or downstream of the coding sequence. In addition, the expression vectors can include additional regulatory sequences from the 3'-untranslated region of plant genes, e.g., a 3' terminator region to increase mRNA stability of the mRNA, such as the PI-II terminator region of potato or the octopine or nopaline synthase 3' terminator regions.

Additional Expression Elements

Specific initiation signals can aid in efficient translation of coding sequences. These signals can include, e.g., the ATG initiation codon and adjacent sequences. In cases where a coding sequence,

its initiation codon and upstream sequences are inserted into the appropriate expression vector, no additional translational control signals may be needed. However, in cases where only coding sequence (e.g., a mature protein coding sequence), or a portion thereof, is inserted, exogenous transcriptional control signals including the ATG initiation codon can be separately provided. The initiation codon is provided in the correct reading frame to facilitate transcription. Exogenous transcriptional elements and initiation codons can be of various origins, both natural and synthetic. The efficiency of expression can be enhanced by the inclusion of enhancers appropriate to the cell system in use.

Expression Hosts

The present invention also relates to host cells which are transduced with vectors of the invention, and the production of polypeptides of the invention (including fragments thereof) by recombinant techniques. Host cells are genetically engineered (i.e., nucleic acids are introduced, e.g., transduced, transformed or transfected) with the vectors of this invention, which may be, for example, a cloning vector or an expression vector comprising the relevant nucleic acids herein. The vector is optionally a plasmid, a viral particle, a phage, a naked nucleic acid, etc. The engineered host cells can be cultured in conventional nutrient media modified as appropriate for activating promoters, selecting transformants, or amplifying the relevant gene. The culture conditions, such as temperature, pH and the like, are those previously used with the host cell selected for expression, and will be apparent to those skilled in the art and in the references cited herein, including, Sambrook, *supra* and Ausubel, *supra*.

The host cell can be a eukaryotic cell, such as a yeast cell, or a plant cell, or the host cell can be a prokaryotic cell, such as a bacterial cell. Plant protoplasts are also suitable for some applications. For example, the DNA fragments are introduced into plant tissues, cultured plant cells or plant protoplasts by standard methods including electroporation (Fromm et al. (1985) *Proc. Natl. Acad. Sci.* 82: 5824-5828, infection by viral vectors such as cauliflower mosaic virus (CaMV) (Hohn et al. (1982) Molecular Biology of Plant Tumors Academic Press, New York, NY, pp. 549-560; US 4,407,956), high velocity ballistic penetration by small particles with the nucleic acid either within the matrix of small beads or particles, or on the surface (Klein et al. (1987) *Nature* 327: 70-73), use of pollen as vector (WO 85/01856), or use of *Agrobacterium tumefaciens* or *A. rhizogenes* carrying a T-DNA plasmid in which DNA fragments are cloned. The T-DNA plasmid is transmitted to plant cells upon infection by *Agrobacterium tumefaciens*, and a portion is stably integrated into the plant genome (Horsch et al. (1984) *Science* 233: 496-498; Fraley et al. (1983) *Proc. Natl. Acad. Sci.* 80: 4803-4807).

The cell can include a nucleic acid of the invention that encodes a polypeptide, wherein the cell expresses a polypeptide of the invention. The cell can also include vector sequences, or the like. Furthermore, cells and transgenic plants that include any polypeptide or nucleic acid above or

throughout this specification, e.g., produced by transduction of a vector of the invention, are an additional feature of the invention.

For long-term, high-yield production of recombinant proteins, stable expression can be used. Host cells transformed with a nucleotide sequence encoding a polypeptide of the invention are optionally cultured under conditions suitable for the expression and recovery of the encoded protein from cell culture. The protein or fragment thereof produced by a recombinant cell may be secreted, membrane-bound, or contained intracellularly, depending on the sequence and/or the vector used. As will be understood by those of skill in the art, expression vectors containing polynucleotides encoding mature proteins of the invention can be designed with signal sequences which direct secretion of the mature polypeptides through a prokaryotic or eukaryotic cell membrane.

Modified Amino Acid Residues

Polypeptides of the invention may contain one or more modified amino acid residues. The presence of modified amino acids may be advantageous in, for example, increasing polypeptide half-life, reducing polypeptide antigenicity or toxicity, increasing polypeptide storage stability, or the like. Amino acid residue(s) are modified, for example, co-translationally or post-translationally during recombinant production or modified by synthetic or chemical means.

Non-limiting examples of a modified amino acid residue include incorporation or other use of acetylated amino acids, glycosylated amino acids, sulfated amino acids, prenylated (e.g., farnesylated, geranylgeranylated) amino acids, PEG modified (e.g., "PEGylated") amino acids, biotinylated amino acids, carboxylated amino acids, phosphorylated amino acids, etc. References adequate to guide one of skill in the modification of amino acid residues are replete throughout the literature.

The modified amino acid residues may prevent or increase affinity of the polypeptide for another molecule, including, but not limited to, polynucleotide, proteins, carbohydrates, lipids and lipid derivatives, and other organic or synthetic compounds.

Identification of Additional Factors

A transcription factor provided by the present invention can also be used to identify additional endogenous or exogenous molecules that can affect a phentoype or trait of interest. On the one hand, such molecules include organic (small or large molecules) and/or inorganic compounds that affect expression of (i.e., regulate) a particular transcription factor. Alternatively, such molecules include endogenous molecules that are acted upon either at a transcriptional level by a transcription factor of the invention to modify a phenotype as desired. For example, the transcription factors can be employed to identify one or more downstream genes that are subject to a regulatory effect of the transcription

factor. In one approach, a transcription factor or transcription factor homolog of the invention is expressed in a host cell, e.g., a transgenic plant cell, tissue or explant, and expression products, either RNA or protein, of likely or random targets are monitored, e.g., by hybridization to a microarray of nucleic acid probes corresponding to genes expressed in a tissue or cell type of interest, by two-dimensional gel electrophoresis of protein products, or by any other method known in the art for assessing expression of gene products at the level of RNA or protein. Alternatively, a transcription factor of the invention can be used to identify promoter sequences (such as binding sites on DNA sequences) involved in the regulation of a downstream target. After identifying a promoter sequence, interactions between the transcription factor and the promoter sequence can be modified by changing specific nucleotides in the promoter sequence or specific amino acids in the transcription factor that interact with the promoter sequence to alter a plant trait. Typically, transcription factor DNA-binding sites are identified by gel shift assays. After identifying the promoter regions, the promoter region sequences can be employed in double-stranded DNA arrays to identify molecules that affect the interactions of the transcription factors with their promoters (Bulyk et al. (1999) *Nature Biotechnol.* 17: 573-577).

The identified transcription factors are also useful to identify proteins that modify the activity of the transcription factor. Such modification can occur by covalent modification, such as by phosphorylation, or by protein-protein (homo or-heteropolymer) interactions. Any method suitable for detecting protein-protein interactions can be employed. Among the methods that can be employed are co-immunoprecipitation, cross-linking and co-purification through gradients or chromatographic columns, and the two-hybrid yeast system.

The two-hybrid system detects protein interactions in vivo and is described in Chien et al. ((1991) *Proc. Natl. Acad. Sci.* 88: 9578-9582) and is commercially available from Clontech (Palo Alto, Calif.). In such a system, plasmids are constructed that encode two hybrid proteins: one consists of the DNA-binding domain of a transcription activator protein fused to the TF polypeptide and the other consists of the transcription activator protein's activation domain fused to an unknown protein that is encoded by a cDNA that has been recombined into the plasmid as part of a cDNA library. The DNA-binding domain fusion plasmid and the cDNA library are transformed into a strain of the yeast *Saccharomyces cerevisiae* that contains a reporter gene (e.g., lacZ) whose regulatory region contains the transcription activator's binding site. Either hybrid protein alone cannot activate transcription of the reporter gene. Interaction of the two hybrid proteins reconstitutes the functional activator protein and results in expression of the reporter gene, which is detected by an assay for the reporter gene product. Then, the library plasmids responsible for reporter gene expression are isolated and sequenced to identify the proteins encoded by the library plasmids. After identifying proteins that interact with the

transcription factors, assays for compounds that interfere with the TF protein-protein interactions can be preformed.

Identification of Modulators

In addition to the intracellular molecules described above, extracellular molecules that alter activity or expression of a transcription factor, either directly or indirectly, can be identified. For example, the methods can entail first placing a candidate molecule in contact with a plant or plant cell. The molecule can be introduced by topical administration, such as spraying or soaking of a plant, or incubating a plant in a solution containing the molecule, and then the molecule's effect on the expression or activity of the TF polypeptide or the expression of the polynucleotide monitored. Changes in the expression of the TF polypeptide can be monitored by use of polyclonal or monoclonal antibodies, gel electrophoresis or the like. Changes in the expression of the corresponding polynucleotide sequence can be detected by use of microarrays, Northerns, quantitative PCR, or any other technique for monitoring changes in mRNA expression. These techniques are exemplified in Ausubel et al. (eds.) Current Protocols in Molecular Biology, John Wiley & Sons (1998, and supplements through 2001). Changes in the activity of the transcription factor can be monitored, directly or indirectly, by assaying the function of the transcription factor, for example, by measuring the expression of promoters known to be controlled by the transcription factor (using promoter-reporter constructs), measuring the levels of transcripts using microarrays, Northern blots, quantitative PCR, etc. Such changes in the expression levels can be correlated with modified plant traits and thus identified molecules can be useful for soaking or spraying on fruit, vegetable and grain crops to modify traits in plants.

Essentially any available composition can be tested for modulatory activity of expression or activity of any nucleic acid or polypeptide herein. Thus, available libraries of compounds such as chemicals, polypeptides, nucleic acids and the like can be tested for modulatory activity. Often, potential modulator compounds can be dissolved in aqueous or organic (e.g., DMSO-based) solutions for easy delivery to the cell or plant of interest in which the activity of the modulator is to be tested. Optionally, the assays are designed to screen large modulator composition libraries by automating the assay steps and providing compounds from any convenient source to assays, which are typically run in parallel (e.g., in microtiter formats on micrometer plates in robotic assays).

In one embodiment, high throughput screening methods involve providing a combinatorial library containing a large number of potential compounds (potential modulator compounds). Such "combinatorial chemical libraries" are then screened in one or more assays, as described herein, to identify those library members (particular chemical species or subclasses) that display a desired

characteristic activity. The compounds thus identified can serve as target compounds.

A combinatorial chemical library can be, e.g., a collection of diverse chemical compounds generated by chemical synthesis or biological synthesis. For example, a combinatorial chemical library such as a polypeptide library is formed by combining a set of chemical building blocks (e.g., in one example, amino acids) in every possible way for a given compound length (i.e., the number of amino acids in a polypeptide compound of a set length). Exemplary libraries include peptide libraries, nucleic acid libraries, antibody libraries (see, e.g., Vaughn et al. (1996) *Nature Biotechnol.* 14: 309-314 and PCT/US96/10287), carbohydrate libraries (see, e.g., Liang et al. *Science* (1996) 274: 1520-1522 and US Patent 5,593,853), peptide nucleic acid libraries (see, e.g., US Patent 5,539,083), and small organic molecule libraries (see, e.g., benzodiazepines, in Baum *Chem. & Engineering News* Jan 18, 1993, page 33; isoprenoids, US Patent 5,569,588; thiazolidinones and metathiazanones, US Patent 5,549,974; pyrrolidines, US Patents 5,525,735 and 5,519,134; morpholino compounds, US Patent 5,506,337) and the like.

Preparation and screening of combinatorial or other libraries is well known to those of skill in the art. Such combinatorial chemical libraries include, but are not limited to, peptide libraries (see, e.g., US Patent 5,010,175; Furka, (1991) *Int. J. Pept. Prot. Res.* 37: 487-493; and Houghton et al. (1991) *Nature* 354: 84-88). Other chemistries for generating chemical diversity libraries can also be used.

In addition, as noted, compound screening equipment for high-throughput screening is generally available, e.g., using any of a number of well known robotic systems that have also been developed for solution phase chemistries useful in assay systems. These systems include automated workstations including an automated synthesis apparatus and robotic systems utilizing robotic arms. Any of the above devices are suitable for use with the present invention, e.g., for high-throughput screening of potential modulators. The nature and implementation of modifications to these devices (if any) so that they can operate as discussed herein will be apparent to persons skilled in the relevant art.

Indeed, entire high-throughput screening systems are commercially available. These systems typically automate entire procedures including all sample and reagent pipetting, liquid dispensing, timed incubations, and final readings of the microplate in detector(s) appropriate for the assay. These configurable systems provide high throughput and rapid start up as well as a high degree of flexibility and customization. Similarly, microfluidic implementations of screening are also commercially available.

The manufacturers of such systems provide detailed protocols the various high throughput. Thus, for example, Zymark Corp. provides technical bulletins describing screening systems for detecting the modulation of gene transcription, ligand binding, and the like. The integrated systems herein, in addition to providing for sequence alignment and, optionally, synthesis of relevant nucleic

acids, can include such screening apparatus to identify modulators that have an effect on one or more polynucleotides or polypeptides according to the present invention.

In some assays it is desirable to have positive controls to ensure that the components of the assays are working properly. At least two types of positive controls are appropriate. That is, known transcriptional activators or inhibitors can be incubated with cells or plants, for example, in one sample of the assay, and the resulting increase/decrease in transcription can be detected by measuring the resulting increase in RNA levels and/or protein expression, for example, according to the methods herein. It will be appreciated that modulators can also be combined with transcriptional activators or inhibitors to find modulators that inhibit transcriptional activation or transcriptional repression. Either expression of the nucleic acids and proteins herein or any additional nucleic acids or proteins activated by the nucleic acids or proteins herein, or both, can be monitored.

In an embodiment, the invention provides a method for identifying compositions that modulate the activity or expression of a polynucleotide or polypeptide of the invention. For example, a test compound, whether a small or large molecule, is placed in contact with a cell, plant (or plant tissue or explant), or composition comprising the polynucleotide or polypeptide of interest and a resulting effect on the cell, plant, (or tissue or explant) or composition is evaluated by monitoring, either directly or indirectly, one or more of: expression level of the polynucleotide or polypeptide, activity (or modulation of the activity) of the polynucleotide or polypeptide. In some cases, an alteration in a plant phenotype can be detected following contact of a plant (or plant cell, or tissue or explant) with the putative modulator, e.g., by modulation of expression or activity of a polynucleotide or polypeptide of the invention. Modulation of expression or activity of a polynucleotide or polypeptide of the invention may also be caused by molecular elements in a signal transduction second messenger pathway and such modulation can affect similar elements in the same or another signal transduction second messenger pathway.

Subsequences

Also contemplated are uses of polynucleotides, also referred to herein as oligonucleotides, typically having at least 12 bases, preferably at least 15, more preferably at least 20, 30, or 50 bases, which hybridize under at least highly stringent (or ultra-high stringent or ultra-ultra-high stringent conditions) conditions to a polynucleotide sequence described above. The polynucleotides may be used as probes, primers, sense and antisense agents, and the like, according to methods as noted *supra*.

Subsequences of the polynucleotides of the invention, including polynucleotide fragments and oligonucleotides are useful as nucleic acid probes and primers. An oligonucleotide suitable for use as a probe or primer is at least about 15 nucleotides in length, more often at least about 18 nucleotides,

often at least about 21 nucleotides, frequently at least about 30 nucleotides, or about 40 nucleotides, or more in length. A nucleic acid probe is useful in hybridization protocols, e.g., to identify additional polypeptide homologs of the invention, including protocols for microarray experiments. Primers can be annealed to a complementary target DNA strand by nucleic acid hybridization to form a hybrid between the primer and the target DNA strand, and then extended along the target DNA strand by a DNA polymerase enzyme. Primer pairs can be used for amplification of a nucleic acid sequence, e.g., by the polymerase chain reaction (PCR) or other nucleic-acid amplification methods. See Sambrook, *supra*, and Ausubel, *supra*.

In addition, the invention includes an isolated or recombinant polypeptide including a subsequence of at least about 15 contiguous amino acids encoded by the recombinant or isolated polynucleotides of the invention. For example, such polypeptides, or domains or fragments thereof, can be used as immunogens, e.g., to produce antibodies specific for the polypeptide sequence, or as probes for detecting a sequence of interest. A subsequence can range in size from about 15 amino acids in length up to and including the full length of the polypeptide.

To be encompassed by the present invention, an expressed polypeptide which comprises such a polypeptide subsequence performs at least one biological function of the intact polypeptide in substantially the same manner, or to a similar extent, as does the intact polypeptide. For example, a polypeptide fragment can comprise a recognizable structural motif or functional domain such as a DNA binding domain that activates transcription, e.g., by binding to a specific DNA promoter region an activation domain, or a domain for protein-protein interactions.

Production of Transgenic Plants

Modification of Traits

The polynucleotides of the invention are favorably employed to produce transgenic plants with various traits, or characteristics, that have been modified in a desirable manner, e.g., to improve the seed characteristics of a plant. For example, alteration of expression levels or patterns (e.g., spatial or temporal expression patterns) of one or more of the transcription factors (or transcription factor homologs) of the invention, as compared with the levels of the same protein found in a wild-type plant, can be used to modify a plant's traits. An illustrative example of trait modification, improved characteristics, by altering expression levels of a particular transcription factor is described further in the Examples and the Sequence Listing.

Arabidopsis as a model system

Arabidopsis thaliana is the object of rapidly growing attention as a model for genetics and

metabolism in plants. *Arabidopsis* has a small genome, and well-documented studies are available. It is easy to grow in large numbers and mutants defining important genetically controlled mechanisms are either available, or can readily be obtained. Various methods to introduce and express isolated homologous genes are available (see Koncz et al., eds., Methods in *Arabidopsis* Research (1992) World Scientific, New Jersey, NJ, in "Preface"). Because of its small size, short life cycle, obligate autogamy and high fertility, *Arabidopsis* is also a choice organism for the isolation of mutants and studies in morphogenetic and development pathways, and control of these pathways by transcription factors (Koncz *supra*, p. 72). A number of studies introducing transcription factors into *A. thaliana* have demonstrated the utility of this plant for understanding the mechanisms of gene regulation and trait alteration in plants. (See, for example, Koncz *supra*, and US Patent Number 6,417,428).

Arabidopsis genes in transgenic plants.

Expression of genes which encode transcription factors modify expression of endogenous genes, polynucleotides, and proteins are well known in the art. In addition, transgenic plants comprising isolated polynucleotides encoding transcription factors may also modify expression of endogenous genes, polynucleotides, and proteins. Examples include Peng et al. (1997 *Genes and Development* 11: 3194-3205) and Peng et al. (1999 *Nature* 400: 256-261). In addition, many others have demonstrated that an *Arabidopsis* transcription factor expressed in an exogenous plant species elicits the same or very similar phenotypic response. See, for example, Fu et al. (2001 *Plant Cell* 13: 1791-1802); Nandi et al. (2000 *Curr. Biol.* 10: 215-218); Coupland (1995 *Nature* 377: 482-483); and Weigel and Nilsson (1995, *Nature* 377: 482-500).

Homologous genes introduced into transgenic plants.

Homologous genes that may be derived from any plant, or from any source whether natural, synthetic, semi-synthetic or recombinant, and that share significant sequence identity or similarity to those provided by the present invention, may be introduced into plants, for example, crop plants, to confer desirable or improved traits. Consequently, transgenic plants may be produced that comprise a recombinant expression vector or cassette with a promoter operably linked to one or more sequences homologous to presently disclosed sequences. The promoter may be, for example, a plant or viral promoter.

The invention thus provides for methods for preparing transgenic plants, and for modifying plant traits. These methods include introducing into a plant a recombinant expression vector or cassette comprising a functional promoter operably linked to one or more sequences homologous to presently disclosed sequences. Plants and kits for producing these plants that result from the application of these

methods are also encompassed by the present invention.

Transcription factors of interest for the modification of plant traits

Currently, the existence of a series of maturity groups for different latitudes represents a major barrier to the introduction of new valuable traits. Any trait (e.g. disease resistance) has to be bred into each of the different maturity groups separately, a laborious and costly exercise. The availability of single strain, which could be grown at any latitude, would therefore greatly increase the potential for introducing new traits to crop species such as soybean and cotton.

For the specific effects, traits and utilities conferred to plants, one or more transcription factor genes of the present invention may be used to increase or decrease, or improve or prove deleterious to a given trait. For example, knocking out a transcription factor gene that naturally occurs in a plant, or suppressing the gene (with, for example, antisense suppression), may cause decreased tolerance to an osmotic stress relative to non-transformed or wild-type plants. By overexpressing this gene, the plant may experience increased tolerance to the same stress. More than one transcription factor gene may be introduced into a plant, either by transforming the plant with one or more vectors comprising two or more transcription factors, or by selective breeding of plants to yield hybrid crosses that comprise more than one introduced transcription factor.

Genes, traits and utilities that affect plant characteristics

Plant transcription factors can modulate gene expression, and, in turn, be modulated by the environmental experience of a plant. Significant alterations in a plant's environment invariably result in a change in the plant's transcription factor gene expression pattern. Altered transcription factor expression patterns generally result in phenotypic changes in the plant. Transcription factor gene product(s) in transgenic plants then differ(s) in amounts or proportions from that found in wild-type or non-transformed plants, and those transcription factors likely represent polypeptides that are used to alter the response to the environmental change. By way of example, it is well accepted in the art that analytical methods based on altered expression patterns may be used to screen for phenotypic changes in a plant far more effectively than can be achieved using traditional methods.

Sugar sensing.

In addition to their important role as an energy source and structural component of the plant cell, sugars are central regulatory molecules that control several aspects of plant physiology, metabolism and development (Hsieh et al. (1998) *Proc. Natl. Acad. Sci.* 95: 13965-13970). It is thought that this control is achieved by regulating gene expression and, in higher plants, sugars have been shown to repress or activate plant genes involved in many essential processes such as photosynthesis, glyoxylate metabolism, respiration, starch and sucrose synthesis and degradation,

pathogen response, wounding response, cell cycle regulation, pigmentation, flowering and senescence. The mechanisms by which sugars control gene expression are not understood.

Several sugar sensing mutants have turned out to be allelic to abscisic acid (ABA) and ethylene mutants. ABA is found in all photosynthetic organisms and acts as a key regulator of transpiration, stress responses, embryogenesis, and seed germination. Most ABA effects are related to the compound acting as a signal of decreased water availability, whereby it triggers a reduction in water loss, slows growth, and mediates adaptive responses. However, ABA also influences plant growth and development via interactions with other phytohormones. Physiological and molecular studies indicate that maize and *Arabidopsis* have almost identical pathways with regard to ABA biosynthesis and signal transduction. For further review, see Finkelstein and Rock ((2002) Abscisic acid biosynthesis and response (In The *Arabidopsis* Book, Editors: Somerville and Meyerowitz (American Society of Plant Biologists, Rockville, MD).

This potentially implicates G481 and G482 in hormone signaling based on the sucrose sugar sensing phenotype of 35S::G481 and 35S::G482 transgenic lines. On the other hand, under the laboratory conditions we use at Mendel, the sucrose treatment (9.5% w/v) could also be an osmotic stress. Therefore, one could interpret this data to indicate that the 35S::G481 transgenic lines are more tolerant to osmotic stress. Interestingly, the Mendel RT-PCR expression profiling studies have shown that more than half of the CCAAT transcription factors are up-regulated in tissues with developing seeds. One example is the well-characterized HAP3-like protein, LEC1, which is required for desiccation tolerance during seed maturation. LEC1 is also ABA and drought inducible. This information, combined with the fact that CCAAT genes are disproportionately responsive to osmotic stress suggests that this family of transcription factors could control pathways involved in both ABA responses and desiccation tolerance.

Because sugars are important signaling molecules, the ability to control either the concentration of a signaling sugar or how the plant perceives or responds to a signaling sugar could be used to control plant development, physiology or metabolism. For example, the flux of sucrose (a disaccharide sugar used for systemically transporting carbon and energy in most plants) has been shown to affect gene expression and alter storage compound accumulation in seeds. Manipulation of the sucrose-signaling pathway in seeds may therefore cause seeds to have more protein, oil or carbohydrate, depending on the type of manipulation. Similarly, in tubers, sucrose is converted to starch which is used as an energy store. It is thought that sugar-signaling pathways may partially determine the levels of starch synthesized in the tubers. The manipulation of sugar signaling in tubers could lead to tubers with a higher starch content.

Thus, the presently disclosed transcription factor genes that manipulate the sugar signal

transduction pathway, including, for example, G481, along with its equivalogs, may lead to altered gene expression to produce plants with desirable traits. In particular, manipulation of sugar signal transduction pathways could be used to alter source-sink relationships in seeds, tubers, roots and other storage organs leading to increase in yield.

Osmotic stress. Modification of the expression of a number of presently disclosed transcription factor genes may be used to increase germination rate or growth under adverse osmotic conditions, which could impact survival and yield of seeds and plants. Osmotic stresses may be regulated by specific molecular control mechanisms that include genes controlling water and ion movements, functional and structural stress-induced proteins, signal perception and transduction, and free radical scavenging, and many others (Wang et al. (2001) *Acta Hort.* (ISHS) 560: 285-292). Instigators of osmotic stress include freezing, drought and high salinity, each of which is discussed in more detail below.

In many ways, freezing, high salt and drought have similar effects on plants, not the least of which is the induction of common polypeptides that respond to these different stresses. For example, freezing is similar to water deficit in that freezing reduces the amount of water available to a plant. Exposure to freezing temperatures may lead to cellular dehydration as water leaves cells and forms ice crystals in intercellular spaces (Buchanan, *supra*). As with high salt concentration and freezing, the problems for plants caused by low water availability include mechanical stresses caused by the withdrawal of cellular water. Thus, the incorporation of transcription factors that modify a plant's response to osmotic stress into, for example, a crop or ornamental plant, may be useful in reducing damage or loss. Specific effects caused by freezing, high salt and drought are addressed below.

Salt and drought tolerance

Plants are subject to a range of environmental challenges. Several of these, including salt stress, general osmotic stress, drought stress and freezing stress, have the ability to impact whole plant and cellular water availability. Not surprisingly, then, plant responses to this collection of stresses are related. In a recent review, Zhu notes that "most studies on water stress signaling have focused on salt stress primarily because plant responses to salt and drought are closely related and the mechanisms overlap" (Zhu (2002) *Ann. Rev. Plant Biol.* 53: 247-273). Many examples of similar responses and pathways to this set of stresses have been documented. For example, the CBF transcription factors have been shown to condition resistance to salt, freezing and drought (Kasuga et al. (1999) *Nature Biotech.* 17: 287-291). The *Arabidopsis rd29B* gene is induced in response to both salt and dehydration stress, a process that is mediated largely through an ABA signal transduction process (Uno et al. (2000) *Proc. Natl. Acad. Sci.* USA 97: 11632-11637), resulting in altered activity of transcription factors that bind to

an upstream element within the *rd29B* promoter. In *Mesembryanthemum crystallinum* (ice plant), Patharker and Cushman have shown that a calcium-dependent protein kinase (McCDPK1) is induced by exposure to both drought and salt stresses (Patharker and Cushman (2000) *Plant J.* 24: 679-691). The stress-induced kinase was also shown to phosphorylate a transcription factor, presumably altering its activity, although transcript levels of the target transcription factor are not altered in response to salt or drought stress. Similarly, Saijo et al. demonstrated that a rice salt/drought-induced calmodulin-dependent protein kinase (OsCDPK7) conferred increased salt and drought tolerance to rice when overexpressed (Saijo et al. (2000) *Plant J.* 23: 319-327).

Exposure to dehydration invokes similar survival strategies in plants as does freezing stress (see, for example, Yelenosky (1989) *Plant Physiol* 89: 444-451) and drought stress induces freezing tolerance (see, for example, Siminovitch et al. (1982) *Plant Physiol* 69: 250-255; and Guy et al. (1992) *Planta* 188: 265-270). In addition to the induction of cold-acclimation proteins, strategies that allow plants to survive in low water conditions may include, for example, reduced surface area, or surface oil or wax production.

Consequently, one skilled in the art would expect that some pathways involved in resistance to one of these stresses, and hence regulated by an individual transcription factor, will also be involved in resistance to another of these stresses, regulated by the same or homologous transcription factors. Of course, the overall resistance pathways are related, not identical, and therefore not all transcription factors controlling resistance to one stress will control resistance to the other stresses. Nonetheless, if a transcription factor conditions resistance to one of these stresses, it would be apparent to one skilled in the art to test for resistance to these related stresses.

Thus, modifying the expression of a number of presently disclosed transcription factor genes, such as G481 or G482, may be used to increase a plant's tolerance to low water conditions and provide the benefits of improved survival, increased yield and an extended geographic and temporal planting range.

Salt. The genes of the sequence listing, including, for example, G482, that provide tolerance to salt may be used to engineer salt tolerant crops and trees that can flourish in soils with high saline content or under drought conditions. In particular, increased salt tolerance during the germination stage of a plant enhances survival and yield. Presently disclosed transcription factor genes that provide increased salt tolerance during germination, the seedling stage, and throughout a plant's life cycle, would find particular value for imparting survival and yield in areas where a particular crop would not normally prosper.

Increased anthocyanin level in plant organs and tissues. Presently disclosed transcription factor genes (i.e., G481 and its equivalogs) can be used to alter anthocyanin levels in one or more tissues,

depending on the organ in which these genes are expressed. The potential utilities of these genes include alterations in pigment production for horticultural purposes, and possibly increasing stress resistance, including osmotic stress resistance. In addition, plants with increased anthocyanin may provide health-promoting effects such as inhibition of tumor growth, prevention of bone loss and prevention of the oxidation of lipids.

Summary of altered plant characteristics. A clade of structurally and functionally related sequences that derive from a wide range of plants, including polynucleotide SEQ ID NOs: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, polynucleotides that encode polypeptide SEQ ID NOs: 29-32, fragments thereof, paralogs, orthologs, equivalogs, and fragments thereof, is provided. These sequences have been shown in laboratory and field experiments to confer altered size and abiotic stress tolerance phenotypes in plants. The invention also provides polypeptides comprising SEQ ID NOs: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24,26, 28, 29, 30, 31, 32, 34, 36, 38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 62, 64, 66, 68, 70, 72, 74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, and fragments thereof, conserved domains thereof, paralogs, orthologs, equivalogs, and fragments thereof. Plants that overexpress these sequences have been observed to be more tolerant to a wide variety of abiotic stresses, including, germination in heat and cold, and osmotic stresses such as drought and high salt levels. Many of the orthologs of these sequences are listed in the Sequence Listing, and due to the high degree of structural similarity to the sequences of the invention, it is expected that these sequences may also function to increase plant biomass and/or abiotic stress tolerance. The invention also encompasses the complements of the polynucleotides. The polynucleotides are useful for screening libraries of molecules or compounds for specific binding and for creating transgenic plants having increased biomass and/or abiotic stress tolerance.

Antisense and Co-suppression

In addition to expression of the nucleic acids of the invention as gene replacement or plant phenotype modification nucleic acids, the nucleic acids are also useful for sense and anti-sense suppression of expression, e.g. to down-regulate expression of a nucleic acid of the invention, e.g. as a further mechanism for modulating plant phenotype. That is, the nucleic acids of the invention, or subsequences or anti-sense sequences thereof, can be used to block expression of naturally occurring homologous nucleic acids. A variety of sense and anti-sense technologies are known in the art, e.g. as set forth in Lichtenstein and Nellen (1997) Antisense Technology: A Practical Approach IRL Press at Oxford University Press, Oxford, U.K. Antisense regulation is also described in Crowley et al. (1985)

Cell 43: 633-641; Rosenberg et al. (1985) *Nature* 313: 703-706; Preiss et al. (1985) *Nature* 313: 27-32; Melton (1985) *Proc. Natl. Acad. Sci.* 82: 144-148; Izant and Weintraub (1985) *Science* 229: 345-352; and Kim and Wold (1985) *Cell* 42: 129-138. Additional methods for antisense regulation are known in the art. Antisense regulation has been used to reduce or inhibit expression of plant genes in, for example in European Patent Publication No. 271988. Antisense RNA may be used to reduce gene expression to produce a visible or biochemical phenotypic change in a plant (Smith et al. (1988) *Nature*, 334: 724-726; Smith et al. (1990) *Plant Mol. Biol.* 14: 369-379). In general, sense or anti-sense sequences are introduced into a cell, where they are optionally amplified, e.g. by transcription. Such sequences include both simple oligonucleotide sequences and catalytic sequences such as ribozymes.

For example, a reduction or elimination of expression (i.e., a "knock-out") of a transcription factor or transcription factor homolog polypeptide in a transgenic plant, e.g., to modify a plant trait, can be obtained by introducing an antisense construct corresponding to the polypeptide of interest as a cDNA. For antisense suppression, the transcription factor or homolog cDNA is arranged in reverse orientation (with respect to the coding sequence) relative to the promoter sequence in the expression vector. The introduced sequence need not be the full-length cDNA or gene, and need not be identical to the cDNA or gene found in the plant type to be transformed. Typically, the antisense sequence need only be capable of hybridizing to the target gene or RNA of interest. Thus, where the introduced sequence is of shorter length, a higher degree of homology to the endogenous transcription factor sequence will be needed for effective antisense suppression. While antisense sequences of various lengths can be utilized, preferably, the introduced antisense sequence in the vector will be at least 30 nucleotides in length, and improved antisense suppression will typically be observed as the length of the antisense sequence increases. Preferably, the length of the antisense sequence in the vector will be greater than 100 nucleotides. Transcription of an antisense construct as described results in the production of RNA molecules that are the reverse complement of mRNA molecules transcribed from the endogenous transcription factor gene in the plant cell.

Suppression of endogenous transcription factor gene expression can also be achieved using RNA interference (RNAi) or microRNA-based methods (Llave et al. (2002) *Science* 297: 2053-2056; Tang et al. (2003) *Genes Dev.* 17: 49-63). RNAi is a post-transcriptional, targeted gene-silencing technique that uses double-stranded RNA (dsRNA) to incite degradation of messenger RNA (mRNA) containing the same sequence as the dsRNA (Constans, (2002) *The Scientist* 16: 36). Small interfering RNAs, or siRNAs are produced in at least two steps: an endogenous ribonuclease cleaves longer dsRNA into shorter, 21-23 nucleotide-long RNAs (Plasterk (2002) *Science* 296: 1263-1265). The siRNA segments then mediate the degradation of the target mRNA (Zamore, (2001) *Nature Struct.*

Biol., 8:746-50). RNAi has been used for gene function determination in a manner similar to antisense oligonucleotides (Constans, (2002) *The Scientist* 16:36). Expression vectors that continually express siRNAs in transiently and stably transfected have been engineered to express small hairpin RNAs (shRNAs), which get processed in vivo into siRNAs-like molecules capable of carrying out gene-specific silencing (Brummelkamp et al., (2002) *Science* 296:550-553, and Paddison, et al. (2002) *Genes & Dev.* 16:948-958). Post-transcriptional gene silencing by double-stranded RNA is discussed in further detail by Hammond et al. (2001) *Nature Rev Gen* 2: 110-119, Fire et al. (1998) *Nature* 391: 806-811 and Timmons and Fire (1998) *Nature* 395: 854. Vectors in which RNA encoded by a transcription factor or transcription factor homolog cDNA is over-expressed can also be used to obtain co-suppression of a corresponding endogenous gene, e.g., in the manner described in US Patent No. 5,231,020 to Jorgensen. Such co-suppression (also termed sense suppression) does not require that the entire transcription factor cDNA be introduced into the plant cells, nor does it require that the introduced sequence be exactly identical to the endogenous transcription factor gene of interest. However, as with antisense suppression, the suppressive efficiency will be enhanced as specificity of hybridization is increased, e.g., as the introduced sequence is lengthened, and/or as the sequence similarity between the introduced sequence and the endogenous transcription factor gene is increased.

Vectors expressing an untranslatable form of the transcription factor mRNA, e.g., sequences comprising one or more stop codon, or nonsense mutation) can also be used to suppress expression of an endogenous transcription factor, thereby reducing or eliminating its activity and modifying one or more traits. Methods for producing such constructs are described in US Patent No. 5,583,021. Preferably, such constructs are made by introducing a premature stop codon into the transcription factor gene. Alternatively, a plant trait can be modified by gene silencing using double-strand RNA (Sharp (1999) *Genes and Development* 13: 139-141). Another method for abolishing the expression of a gene is by insertion mutagenesis using the T-DNA of *Agrobacterium tumefaciens*. After generating the insertion mutants, the mutants can be screened to identify those containing the insertion in a transcription factor or transcription factor homolog gene. Plants containing a single transgene insertion event at the desired gene can be crossed to generate homozygous plants for the mutation. Such methods are well known to those of skill in the art (See for example Koncz et al. (1992) Methods in *Arabidopsis* Research, World Scientific Publishing Co. Pte. Ltd., River Edge, NJ).

Alternatively, a plant phenotype can be altered by eliminating an endogenous gene, such as a transcription factor or transcription factor homolog, e.g., by homologous recombination (Kempin et al. (1997) *Nature* 389: 802-803).

A plant trait can also be modified by using the Cre-lox system (for example, as described in US Pat. No. 5,658,772). A plant genome can be modified to include first and second lox sites that are

then contacted with a Cre recombinase. If the lox sites are in the same orientation, the intervening DNA sequence between the two sites is excised. If the lox sites are in the opposite orientation, the intervening sequence is inverted.

The polynucleotides and polypeptides of this invention can also be expressed in a plant in the absence of an expression cassette by manipulating the activity or expression level of the endogenous gene by other means, such as, for example, by ectopically expressing a gene by T-DNA activation tagging (Ichikawa et al. (1997) *Nature* 390 698-701; Kakimoto et al. (1996) *Science* 274: 982-985). This method entails transforming a plant with a gene tag containing multiple transcriptional enhancers and once the tag has inserted into the genome, expression of a flanking gene coding sequence becomes deregulated. In another example, the transcriptional machinery in a plant can be modified so as to increase transcription levels of a polynucleotide of the invention (*See*, e.g., PCT Publications WO 96/06166 and WO 98/53057 which describe the modification of the DNA-binding specificity of zinc finger proteins by changing particular amino acids in the DNA-binding motif).

The transgenic plant can also include the machinery necessary for expressing or altering the activity of a polypeptide encoded by an endogenous gene, for example, by altering the phosphorylation state of the polypeptide to maintain it in an activated state.

Transgenic plants (or plant cells, or plant explants, or plant tissues) incorporating the polynucleotides of the invention and/or expressing the polypeptides of the invention can be produced by a variety of well established techniques as described above. Following construction of a vector, most typically an expression cassette, including a polynucleotide, e.g., encoding a transcription factor or transcription factor homolog, of the invention, standard techniques can be used to introduce the polynucleotide into a plant, a plant cell, a plant explant or a plant tissue of interest. Optionally, the plant cell, explant or tissue can be regenerated to produce a transgenic plant.

The plant can be any higher plant, including gymnosperms, monocotyledonous and dicotyledenous plants. Suitable protocols are available for *Leguminosae* (alfalfa, soybean, clover, etc.), *Umbelliferae* (carrot, celery, parsnip), *Cruciferae* (cabbage, radish, rapeseed, broccoli, etc.), *Curcurbitaceae* (melons and cucumber), *Gramineae* (wheat, corn, rice, barley, millet, etc.), *Solanaceae* (potato, tomato, tobacco, peppers, etc.), and various other crops. See protocols described in Ammirato et al., eds., (1984) Handbook of Plant Cell Culture –Crop Species, Macmillan Publ. Co., New York, NY; Shimamoto et al. (1989) *Nature* 338: 274-276; Fromm et al. (1990) *Bio/Technol.* 8: 833-839; and Vasil et al. (1990) *Bio/Technol.* 8: 429-434.

Transformation and regeneration of both monocotyledonous and dicotyledonous plant cells are now routine, and the selection of the most appropriate transformation technique will be determined by the practitioner. The choice of method will vary with the type of plant to be transformed; those skilled

in the art will recognize the suitability of particular methods for given plant types. Suitable methods can include, but are not limited to: electroporation of plant protoplasts; liposome-mediated transformation; polyethylene glycol (PEG) mediated transformation; transformation using viruses; micro-injection of plant cells; micro-projectile bombardment of plant cells; vacuum infiltration; and *Agrobacterium tumefaciens* mediated transformation. Transformation means introducing a nucleotide sequence into a plant in a manner to cause stable or transient expression of the sequence.

Successful examples of the modification of plant characteristics by transformation with cloned sequences which serve to illustrate the current knowledge in this field of technology, and which are herein incorporated by reference, include: US Patent Nos. 5,571,706; 5,677,175; 5,510,471; 5,750,386; 5,597,945; 5,589,615; 5,750,871; 5,268,526; 5,780,708; 5,538,880; 5,773,269; 5,736,369 and 5,610,042.

Following transformation, plants are preferably selected using a dominant selectable marker incorporated into the transformation vector. Typically, such a marker will confer antibiotic or herbicide resistance on the transformed plants, and selection of transformants can be accomplished by exposing the plants to appropriate concentrations of the antibiotic or herbicide.

After transformed plants are selected and grown to maturity, those plants showing a modified trait are identified. The modified trait can be any of those traits described above. Additionally, to confirm that the modified trait is due to changes in expression levels or activity of the polypeptide or polynucleotide of the invention can be determined by analyzing mRNA expression using Northern blots, RT-PCR or microarrays, or protein expression using immunoblots or Western blots or gel shift assays.

Integrated Systems – Sequence Identity

Additionally, the present invention may be an integrated system, computer or computer readable medium that comprises an instruction set for determining the identity of one or more sequences in a database. In addition, the instruction set can be used to generate or identify sequences that meet any specified criteria. Furthermore, the instruction set may be used to associate or link certain functional benefits, such improved characteristics, with one or more identified sequence.

For example, the instruction set can include, e.g., a sequence comparison or other alignment program, e.g., an available program such as, for example, the Wisconsin Package Version 10.0, such as BLAST, FASTA, PILEUP, FINDPATTERNS or the like (GCG, Madison, WI). Public sequence databases such as GenBank, EMBL, Swiss-Prot and PIR or private sequence databases such as PHYTOSEQ sequence database (Incyte Genomics, Palo Alto, CA) can be searched.

Alignment of sequences for comparison can be conducted by the local homology algorithm of

Smith and Waterman (1981) *Adv. Appl. Math.* 2: 482-489, by the homology alignment algorithm of Needleman and Wunsch (1970) *J. Mol. Biol.* 48: 443-453, by the search for similarity method of Pearson and Lipman (1988) *Proc. Natl. Acad. Sci.* 85: 2444-2448, by computerized implementations of these algorithms. After alignment, sequence comparisons between two (or more) polynucleotides or polypeptides are typically performed by comparing sequences of the two sequences over a comparison window to identify and compare local regions of sequence similarity. The comparison window can be a segment of at least about 20 contiguous positions, usually about 50 to about 200, more usually about 100 to about 150 contiguous positions. A description of the method is provided in Ausubel et al. *supra*.

A variety of methods for determining sequence relationships can be used, including manual alignment and computer assisted sequence alignment and analysis. This later approach is a preferred approach in the present invention, due to the increased throughput afforded by computer assisted methods. As noted above, a variety of computer programs for performing sequence alignment are available, or can be produced by one of skill.

One example algorithm that is suitable for determining percent sequence identity and sequence similarity is the BLAST algorithm, which is described in Altschul et al. (1990) *J. Mol. Biol.* 215: 403-410. Software for performing BLAST analyses is publicly available, e.g., through the National Library of Medicine's National Center for Biotechnology Information (ncbi.nlm.nih; see at world wide web (www) National Institutes of Health US government (gov) website). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul et al. *supra*). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always > 0) and N (penalty score for mismatching residues; always < 0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring

matrix (see Henikoff and Henikoff (1992) *Proc. Natl. Acad. Sci.* 89: 10915-10919). Unless otherwise indicated, "sequence identity" here refers to the % sequence identity generated from a tblastx using the NCBI version of the algorithm at the default settings using gapped alignments with the filter "off" (see, for example, NIH NLM NCBI website at ncbi.nlm.nih, *supra*).

In addition to calculating percent sequence identity, the BLAST algorithm also performs a statistical analysis of the similarity between two sequences (see, e.g. Karlin and Altschul (1993) *Proc. Natl. Acad. Sci.* 90: 5873-5787). One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a reference sequence (and, therefore, in this context, homologous) if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.1, or less than about 0.01, and or even less than about 0.001. An additional example of a useful sequence alignment algorithm is PILEUP. PILEUP creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments. The program can align, e.g., up to 300 sequences of a maximum length of 5,000 letters.

The integrated system, or computer typically includes a user input interface allowing a user to selectively view one or more sequence records corresponding to the one or more character strings, as well as an instruction set which aligns the one or more character strings with each other or with an additional character string to identify one or more region of sequence similarity. The system may include a link of one or more character strings with a particular phenotype or gene function. Typically, the system includes a user readable output element that displays an alignment produced by the alignment instruction set.

The methods of this invention can be implemented in a localized or distributed computing environment. In a distributed environment, the methods may implemented on a single computer comprising multiple processors or on a multiplicity of computers. The computers can be linked, e.g. through a common bus, but more preferably the computer(s) are nodes on a network. The network can be a generalized or a dedicated local or wide-area network and, in certain preferred embodiments, the computers may be components of an intra-net or an internet.

Thus, the invention provides methods for identifying a sequence similar or homologous to one or more polynucleotides as noted herein, or one or more target polypeptides encoded by the polynucleotides, or otherwise noted herein and may include linking or associating a given plant phenotype or gene function with a sequence. In the methods, a sequence database is provided (locally or across an inter or intra net) and a query is made against the sequence database using the relevant sequences herein and associated plant phenotypes or gene functions.

Any sequence herein can be entered into the database, before or after querying the database. This provides for both expansion of the database and, if done before the querying step, for insertion of control sequences into the database. The control sequences can be detected by the query to ensure the general integrity of both the database and the query. As noted, the query can be performed using a web browser based interface. For example, the database can be a centralized public database such as those noted herein, and the querying can be done from a remote terminal or computer across an internet or intranet.

Any sequence herein can be used to identify a similar, homologous, paralogous, or orthologous sequence in another plant. This provides means for identifying endogenous sequences in other plants that may be useful to alter a trait of progeny plants, which results from crossing two plants of different strain. For example, sequences that encode an ortholog of any of the sequences herein that naturally occur in a plant with a desired trait can be identified using the sequences disclosed herein. The plant is then crossed with a second plant of the same species but which does not have the desired trait to produce progeny which can then be used in further crossing experiments to produce the desired trait in the second plant. Therefore the resulting progeny plant contains no transgenes; expression of the endogenous sequence may also be regulated by treatment with a particular chemical or other means, such as EMR. Some examples of such compounds well known in the art include: ethylene; cytokinins; phenolic compounds, which stimulate the transcription of the genes needed for infection; specific monosaccharides and acidic environments which potentiate vir gene induction; acidic polysaccharides which induce one or more chromosomal genes; and opines; other mechanisms include light or dark treatment (for a review of examples of such treatments, see, Winans (1992) *Microbiol. Rev.* 56: 12-31; Eyal et al. (1992) *Plant Mol. Biol.* 19: 589-599; Chrispeels et al. (2000) *Plant Mol. Biol.* 42: 279-290; Piazza et al. (2002) *Plant Physiol.* 128: 1077-1086).

Table 5 lists sequences discovered to be orthologous to a number of representative transcription factors of the present invention. The column headings include the transcription factors listed by (a) the SEQ ID NO: of the ortholog or nucleotide encoding the ortholog; (b) the GID sequence identifier; (c) the Sequence Identifier or GenBank Accession Number;(d) the species from which the orthologs to the transcription factors are derived; (e) the smallest sum probability relationship to G482 determined by BLAST analysis; and (f) the percent identity of the B domain of the sequence to the same domain in G482.

Table 5. Paralogs and Orthologs and Other Related Genes of Representative *Arabidopsis* Transcription Factor Genes identified using BLAST

SEQ ID NO: of Ortholog or Nucleotide Encoding Ortholog	GID No.	Sequence Identifier or Accession Number	Species from Which Ortholog is Derived	Smallest Sum Probability to G482	Percent Identity of B domain to B domain of G482
1	G481		*Arabidopsis thaliana*		83%
3	G482		*Arabidopsis thaliana*	0.0	100%
5	G485		*Arabidopsis thaliana*		94%
7	G1364		*Arabidopsis thaliana*		85%
9	G2345		*Arabidopsis thaliana*		85%
11		GLYMA-28NOV01-CLUSTER24839_1	*Glycine max*	5E-29	84%
13		GLYMA-28NOV01-CLUSTER31103_1	*Glycine max*	2E-31	85%
15		GLYMA-28NOV01-CLUSTER33504_1	*Glycine max*	1E-41	91%
17	G3476	GLYMA-28NOV01-CLUSTER33504_3	*Glycine max*	3E-58	94%
19	G3475	GLYMA-28NOV01-CLUSTER33504_5	*Glycine max*	6E-58	95%
21		GLYMA-28NOV01-CLUSTER33504_6	*Glycine max*	6E-45	92%
23	G3471	GLYMA-28NOV01-CLUSTER4778_1	*Glycine max*	9E-57	92%
81	G3472		*Glycine max*	9E-57	92%
25	G3470	GLYMA-28NOV01-CLUSTER4778_3	*Glycine max*	8E-9	85%
87	G3394	ORYSA-22JAN02-CLUSTER26105_1	*Oryza sativa*	3E-18	86%
73	G3395		*Oryza sativa*	1E-44	83%
29		OSC12630.C1.p5.fg	*Oryza sativa*	2E-55	90%
30		OSC1404.C1.p3.fg	*Oryza sativa*	4E-39	75%
31		OSC30077.C1.p6.fg	*Oryza sativa*	3E-50	86%
32		OSC5489.C1.p2.fg	*Oryza sativa*	8E-44	83%
60	G3398		*Oryza sativa*	2E-57	90%
33		LIB3732-044-Q6-K6-C4	*Zea mays*	2E-23	87%
35		ZEAMA-08NOV01-CLUSTER719_1	*Zea mays*	7E-19	86%
37		ZEAMA-08NOV01-CLUSTER719_10	*Zea mays*	7E-11	86%
39		ZEAMA-08NOV01-CLUSTER719_2	*Zea mays*	6E-19	86%
41		ZEAMA-08NOV01-CLUSTER719_3	*Zea mays*	6E-7	80%
43		ZEAMA-08NOV01-CLUSTER719_4	*Zea mays*	8E-17	86%
45		ZEAMA-08NOV01-CLUSTER719_5	*Zea mays*	4E-17	86%
47		ZEAMA-08NOV01-	*Zea mays*	5E-23	93%

		CLUSTER90408_1			
49	G3436	ZEAMA-08NOV01-CLUSTER90408_2	*Zea mays*	7E-55	93%
77	G3434		*Zea mays*	2E-44	86%
79	G3435		*Zea mays*	1E-58	93%
51	G3473	GLYMA-28NOV01-CLUSTER33504_4	*Glycine max*	7E-17	83%
83	G3474		*Glycine max*	6E-57	91%
85	G3477		*Glycine max*	5E-47	85%
87	G3478		*Glycine max*	4E-58	95%
53		ORYSA-22JAN02-CLUSTER119015_1	*Oryza sativa*	9E-21	83%
55		Zm_S11418173	*Zea mays*	3E-17	86%
57		Zm_S11434692	*Zea mays*	1E-19	85%
59		Ta_S45374	*Triticum aestivum*	2E-24	85%
61		Ta_S50443	*Triticum aestivum*	9E-24	90%
63		SGN-UNIGENE-46859	*Lycopersicon esculentum*	2E-6	87%
65		SGN-UNIGENE-47447	*Lycopersicon esculentum*	3E-11	91%
		BU238020	*Descurainia sophia*	1.00E-70	
		BG440251	*Gossypium arboreum*	3.00E-56	
		CB290513	*Citrus sinensis*	3.00E-55	
		BF071234	*Glycine max*	1.00E-54	
		BQ799965	*Vitis vinifera*	3.00E-54	
		AX584261	*Eucalyptus grandis*	5.00E-54	
		AX584259	*Momordica charantia*	7.00E-54	
		CD848631	*Helianthus annuus*	6.00E-53	
		BQ488908	*Beta vulgaris*	6.00E-53	
		CD573484	*Zea mays*	8.00E-53	
		gi115840	*Zea mays*	2.40E-51	86%
		gi30409461	*Oryza sativa (japonica* cultivar-group*)*	3.50E-50	86%
		AP004366	*Oryza sativa*	3E-50	
		AC120529	*Oryza sativa (japonica* cultivar-group*)*	7E-46	
		gi15408794	*Oryza sativa*	8.70E-38	75%
		AC108500	*Oryza sativa*	2E-20	
		CD574709	*Poncirus trifoliata*	8.00E-60	
		BQ505706	*Solanum tuberosum*	9.00E-59	
		AC122165	*Medicago truncatula*	9.00E-57	
		AC120529	*Oryza sativa (japonica* cultivar-group*)*	6E-56	
		BQ104671	*Rosa hybrid* cultivar	3.00E-55	
		AX584271	*Glycine max*	6.00E-55	
		AX584265	*Zea mays*	1.00E-54	
		AAAA01003638	*Oryza sativa (indica* cultivar-group*)*	2.00E-54	
		AP005193	*Oryza sativa (japonica* cultivar-group*)*	2.00E-54	
		BU880488	*Populus balsamifera* subsp. *trichocarpa*	2.00E-53	
		BJ248969	*Triticum aestivum*	3.00E-53	
		gi115840	*Zea mays*	1.80E-46	86%

		gi30409461	*Oryza sativa (japonica* cultivar-group)	8.80E-45	86%
		AP004366	*Oryza sativa*	4E-44	
		gi15408794	*Oryza sativa*	1.80E-37	75%
		AP005193	*Oryza sativa* (japonica cultivar-group)	9E-21	
		AC108500	*Oryza sativa*	5E-15	
		CD574709	*Poncirus trifoliata*	9.00E-62	
		BQ505706	*Solanum tuberosum*	4.00E-60	
		BQ996905	*Lactuca sativa*	2.00E-58	
		AAAA01003638	*Oryza sativa (indica* cultivar-group)	3.00E-57	
		AP005193	*Oryza sativa (japonica* cultivar-group)	3.00E-57	
		BQ592365	*Beta vulgaris*	9.00E-57	
		CD438068	*Zea mays*	9.00E-57	
		AX288144	*Physcomitrella patens*	3.00E-56	
		BU880488	*Populus balsamifera* subsp. *trichocarpa*	1.00E-55	
		AX584277	*Glycine max*	6.00E-55	
		gi30409461	*Oryza sativa (japonica* cultivar-group)	4.60E-48	86%
		gi30349365	*Oryza sativa (indica* cultivar-group)	1.10E-39	
		gi15408794	*Oryza sativa*	1.60E-38	75%
		CD823119	*Brassica napus*	1.00E-64	
		BG642751	*Lycopersicon esculentum*	2.00E-60	
		BQ629472	*Glycine max*	6.00E-60	
		BQ405785	*Gossypium arboreum*	6.00E-60	
		BQ488908	*Beta vulgaris*	1.00E-59	
		AX584261	*Eucalyptus grandis*	3.00E-59	
		BQ799965	*Vitis vinifera*	6.00E-59	
		CB290513	*Citrus sinensis*	3.00E-58	
		CD848631	*Helianthus annuus*	3.00E-58	
		CF069249	*Medicago truncatula*	2.00E-57	
		gi115840	*Zea mays*	2.10E-50	86%
		gi30409461	*Oryza sativa (japonica* cultivar-group)	9.50E-48	82%
		CD823119	*Brassica napus*	2.00E-75	
		BG445358	*Gossypium arboreum*	1.00E-64	
		BG642751	*Lycopersicon esculentum*	2.00E-64	
		BQ629472	*Glycine max*	5.00E-63	
		BQ488908	*Beta vulgaris*	6.00E-63	
		AX584261	*Eucalyptus grandis*	7.00E-62	
		BQ799965	*Vitis vinifera*	1.00E-61	
		CD848631	*Helianthus annuus*	2.00E-61	
		CF069249	*Medicago truncatula*	6.00E-61	
		BG599785	*Solanum tuberosum*	7.00E-61	82%
		gi115840	*Zea mays*	6.80E-54	86%
		gi30409459	*Oryza sativa (japonica* cultivar-group)	1.00E-50	83%

Molecular Modeling

Another means that may be used to confirm the utility and function of transcription factor sequences that are orthologous or paralogous to presently disclosed transcription factors is through the use of molecular modeling software. Molecular modeling is routinely used to predict polypeptide structure, and a variety of protein structure modeling programs, such as "Insight II" (Accelrys, Inc.) are commercially available for this purpose. Modeling can thus be used to predict which residues of a polypeptide can be changed without altering function (Crameri et al. (2003) U.S. Patent No. 6, 521, 453). Thus, polypeptides that are sequentially similar can be shown to have a high likelihood of similar function by their structural similarity, which may, for example, be established by comparison of regions of superstructure. The relative tendencies of amino acids to form regions of superstructure (for example, helixes and _-sheets) are well established. For example, O'Neil et al. ((1990) *Science* 250: 646-651) have discussed in detail the helix forming tendencies of amino acids. Tables of relative structure forming activity for amino acids can be used as substitution tables to predict which residues can be functionally substituted in a given region, for example, in DNA-binding domains of known transcription factors and equivalogs. Homologs that are likely to be functionally similar can then be identified.

Of particular interest is the structure of a transcription factor in the region of its conserved domains, such as those identified in Table 1. Structural analyses may be performed by comparing the structure of the known transcription factor around its conserved domain with those of orthologs and paralogs. Analysis of a number of polypeptides within a transcription factor group or clade, including the functionally or sequentially similar polypeptides provided in the Sequence Listing, may also provide an understanding of structural elements required to regulate transcription within a given family.

EXAMPLES

The invention, now being generally described, will be more readily understood by reference to the following examples, which are included merely for purposes of illustration of certain aspects and embodiments of the present invention and are not intended to limit the invention. It will be recognized by one of skill in the art that a transcription factor that is associated with a particular first trait may also be associated with at least one other, unrelated and inherent second trait which was not predicted by the first trait.

The complete descriptions of the traits associated with each polynucleotide of the invention are fully disclosed in Example VIII. The complete description of the transcription factor gene family and

identified B domains of the polypeptide encoded by the polynucleotide is fully disclosed in Table 1.

Example I: Full Length Gene Identification and Cloning

Putative transcription factor sequences (genomic or ESTs) related to known transcription factors were identified in the *Arabidopsis thaliana* GenBank database using the tblastn sequence analysis program using default parameters and a P-value cutoff threshold of –4 or –5 or lower, depending on the length of the query sequence. Putative transcription factor sequence hits were then screened to identify those containing particular sequence strings. If the sequence hits contained such sequence strings, the sequences were confirmed as transcription factors.

Alternatively, *Arabidopsis thaliana* cDNA libraries derived from different tissues or treatments, or genomic libraries were screened to identify novel members of a transcription family using a low stringency hybridization approach. Probes were synthesized using gene specific primers in a standard PCR reaction (annealing temperature 60° C) and labeled with ^{32}P dCTP using the High Prime DNA Labeling Kit (Boehringer Mannheim Corp. (now Roche Diagnostics Corp., Indianapolis, IN). Purified radiolabelled probes were added to filters immersed in Church hybridization medium (0.5 M $NaPO_4$ pH 7.0, 7% SDS, 1% w/v bovine serum albumin) and hybridized overnight at 60°C with shaking. Filters were washed two times for 45 to 60 minutes with 1xSCC, 1% SDS at 60° C.

To identify additional sequence 5' or 3' of a partial cDNA sequence in a cDNA library, 5' and 3' rapid amplification of cDNA ends (RACE) was performed using the MARATHON cDNA amplification kit (Clontech, Palo Alto, CA). Generally, the method entailed first isolating poly(A) mRNA, performing first and second strand cDNA synthesis to generate double stranded cDNA, blunting cDNA ends, followed by ligation of the MARATHON Adaptor to the cDNA to form a library of adaptor-ligated ds cDNA.

Gene-specific primers were designed to be used along with adaptor specific primers for both 5' and 3' RACE reactions. Nested primers, rather than single primers, were used to increase PCR specificity. Using 5' and 3' RACE reactions, 5' and 3' RACE fragments were obtained, sequenced and cloned. The process can be repeated until 5' and 3' ends of the full-length gene were identified. Then the full-length cDNA was generated by PCR using primers specific to 5' and 3' ends of the gene by end-to-end PCR.

Example II: Construction of Expression Vectors

The sequence was amplified from a genomic or cDNA library using primers specific to sequences upstream and downstream of the coding region. The expression vector was pMEN20 or pMEN65, which are both derived from pMON316 (Sanders et al. (1987) *Nucleic Acids Res.* 15:1543-

1558) and contain the CaMV 35S promoter to express transgenes. To clone the sequence into the vector, both pMEN20 and the amplified DNA fragment were digested separately with SalI and NotI restriction enzymes at 37° C for 2 hours. The digestion products were subject to electrophoresis in a 0.8% agarose gel and visualized by ethidium bromide staining. The DNA fragments containing the sequence and the linearized plasmid were excised and purified by using a QIAQUICK gel extraction kit (Qiagen, Valencia CA). The fragments of interest were ligated at a ratio of 3:1 (vector to insert). Ligation reactions using T4 DNA ligase (New England Biolabs, Beverly MA) were carried out at 16° C for 16 hours. The ligated DNAs were transformed into competent cells of the *E. coli* strain DH5alpha by using the heat shock method. The transformations were plated on LB plates containing 50 mg/l kanamycin (Sigma Chemical Co. St. Louis MO). Individual colonies were grown overnight in five milliliters of LB broth containing 50 mg/l kanamycin at 37° C. Plasmid DNA was purified by using Qiaquick Mini Prep kits (Qiagen).

Example III: Transformation of *Agrobacterium* with the Expression Vector

After the plasmid vector containing the gene was constructed, the vector was used to transform *Agrobacterium tumefaciens* cells expressing the gene products. The stock of *Agrobacterium tumefaciens* cells for transformation was made as described by Nagel et al. (1990) *FEMS Microbiol Letts.* 67: 325-328. *Agrobacterium* strain ABI was grown in 250 ml LB medium (Sigma) overnight at 28°C with shaking until an absorbance over 1 cm at 600 nm (A_{600}) of 0.5 – 1.0 was reached. Cells were harvested by centrifugation at 4,000 x g for 15 min at 4° C. Cells were then resuspended in 250 µl chilled buffer (1 mM HEPES, pH adjusted to 7.0 with KOH). Cells were centrifuged again as described above and resuspended in 125 µl chilled buffer. Cells were then centrifuged and resuspended two more times in the same HEPES buffer as described above at a volume of 100 µl and 750 µl, respectively. Resuspended cells were then distributed into 40 µl aliquots, quickly frozen in liquid nitrogen, and stored at -80° C.

Agrobacterium cells were transformed with plasmids prepared as described above following the protocol described by Nagel et al. (*supra*). For each DNA construct to be transformed, 50 – 100 ng DNA (generally resuspended in 10 mM Tris-HCl, 1 mM EDTA, pH 8.0) was mixed with 40 µl of *Agrobacterium* cells. The DNA/cell mixture was then transferred to a chilled cuvette with a 2mm electrode gap and subject to a 2.5 kV charge dissipated at 25 µF and 200 µF using a Gene Pulser II apparatus (Bio-Rad, Hercules, CA). After electroporation, cells were immediately resuspended in 1.0 ml LB and allowed to recover without antibiotic selection for 2 – 4 hours at 28° C in a shaking incubator. After recovery, cells were plated onto selective medium of LB broth containing 100 µg/ml spectinomycin (Sigma) and incubated for 24-48 hours at 28° C. Single colonies were then picked and

inoculated in fresh medium. The presence of the plasmid construct was verified by PCR amplification and sequence analysis.

Example IV: Transformation of *Arabidopsis* Plants with *Agrobacterium tumefaciens* with Expression Vector

After transformation of *Agrobacterium tumefaciens* with plasmid vectors containing the gene, single *Agrobacterium* colonies were identified, propagated, and used to transform *Arabidopsis* plants. Briefly, 500 ml cultures of LB medium containing 50 mg/l kanamycin were inoculated with the colonies and grown at 28° C with shaking for 2 days until an optical absorbance at 600 nm wavelength over 1 cm (A_{600}) of > 2.0 is reached. Cells were then harvested by centrifugation at 4,000 x g for 10 min, and resuspended in infiltration medium (1/2 X Murashige and Skoog salts (Sigma), 1 X Gamborg's B-5 vitamins (Sigma), 5.0% (w/v) sucrose (Sigma), 0.044 μM benzylamino purine (Sigma), 200 μl/l Silwet L-77 (Lehle Seeds) until an A_{600} of 0.8 was reached.

Prior to transformation, *Arabidopsis thaliana* seeds (ecotype Columbia) were sown at a density of ~10 plants per 4" pot onto Pro-Mix BX potting medium (Hummert International) covered with fiberglass mesh (18 mm X 16 mm). Plants were grown under continuous illumination (50-75 $\mu E/m^2/sec$) at 22-23° C with 65-70% relative humidity. After about 4 weeks, primary inflorescence stems (bolts) are cut off to encourage growth of multiple secondary bolts. After flowering of the mature secondary bolts, plants were prepared for transformation by removal of all siliques and opened flowers.

The pots were then immersed upside down in the mixture of *Agrobacterium* infiltration medium as described above for 30 sec, and placed on their sides to allow draining into a 1' x 2' flat surface covered with plastic wrap. After 24 h, the plastic wrap was removed and pots are turned upright. The immersion procedure was repeated one week later, for a total of two immersions per pot. Seeds were then collected from each transformation pot and analyzed following the protocol described below.

Example V: Identification of *Arabidopsis* Primary Transformants

Seeds collected from the transformation pots were sterilized essentially as follows. Seeds were dispersed into in a solution containing 0.1% (v/v) Triton X-100 (Sigma) and sterile water and washed by shaking the suspension for 20 min. The wash solution was then drained and replaced with fresh wash solution to wash the seeds for 20 min with shaking. After removal of the ethanol/detergent solution, a solution containing 0.1% (v/v) Triton X-100 and 30% (v/v) bleach (CLOROX; Clorox Corp. Oakland CA) was added to the seeds, and the suspension was shaken for 10 min. After removal of the bleach/detergent solution, seeds were then washed five times in sterile distilled water. The seeds

were stored in the last wash water at 4° C for 2 days in the dark before being plated onto antibiotic selection medium (1 X Murashige and Skoog salts (pH adjusted to 5.7 with 1M KOH), 1 X Gamborg's B-5 vitamins, 0.9% phytagar (Life Technologies), and 50 mg/l kanamycin). Seeds were germinated under continuous illumination (50-75 $\mu E/m^2/sec$) at 22-23° C. After 7-10 days of growth under these conditions, kanamycin resistant primary transformants (T1 generation) were visible and obtained. These seedlings were transferred first to fresh selection plates where the seedlings continued to grow for 3-5 more days, and then to soil (Pro-Mix BX potting medium).

Primary transformants were crossed and progeny seeds (T_2) collected; kanamycin resistant seedlings were selected and analyzed. The expression levels of the recombinant polynucleotides in the transformants vary from about a 5% expression level increase to a least a 100% expression level increase. Similar observations are made with respect to polypeptide level expression.

Example VI: Identification of *Arabidopsis* Plants with Transcription Factor Gene Knockouts

The screening of insertion mutagenized *Arabidopsis* collections for null mutants in a known target gene was essentially as described in Krysan et al. (1999) *Plant Cell* 11: 2283-2290. Briefly, gene-specific primers, nested by 5-250 base pairs to each other, were designed from the 5' and 3' regions of a known target gene. Similarly, nested sets of primers were also created specific to each of the T-DNA or transposon ends (the "right" and "left" borders). All possible combinations of gene specific and T-DNA/transposon primers were used to detect by PCR an insertion event within or close to the target gene. The amplified DNA fragments were then sequenced which allows the precise determination of the T-DNA/transposon insertion point relative to the target gene. Insertion events within the coding or intervening sequence of the genes were deconvoluted from a pool comprising a plurality of insertion events to a single unique mutant plant for functional characterization. The method is described in more detail in Yu and Adam, US Application Serial No. 09/177,733 filed October 23, 1998.

Example VII: Identification of Modified Phenotypes in Overexpression or Gene Knockout Plants.

Experiments were performed to identify those transformants or knockouts that exhibited modified biochemical characteristics..

Calibration of NIRS response was performed using data obtained by wet chemical analysis of a population of *Arabidopsis* ecotypes that were expected to represent diversity of oil and protein levels.

Experiments were performed to identify those transformants or knockouts that exhibited modified sugar-sensing. For such studies, seeds from transformants were germinated on media containing 5% glucose or 9.4% sucrose which normally partially restrict hypocotyl elongation. Plants

with altered sugar sensing may have either longer or shorter hypocotyls than normal plants when grown on this media. Additionally, other plant traits may be varied such as root mass.

In some instances, expression patterns of the stress-induced genes may be monitored by microarray experiments. In these experiments, cDNAs are generated by PCR and resuspended at a final concentration of ~ 100 ng/ul in 3X SSC or 150mM Na-phosphate (Eisen and Brown (1999) *Methods Enzymol.* 303: 179-205). The cDNAs are spotted on microscope glass slides coated with polylysine. The prepared cDNAs are aliquoted into 384 well plates and spotted on the slides using, for example, an x-y-z gantry (OmniGrid) which may be purchased from GeneMachines (Menlo Park, CA) outfitted with quill type pins which may be purchased from Telechem International (Sunnyvale, CA). After spotting, the arrays are cured for a minimum of one week at room temperature, rehydrated and blocked following the protocol recommended by Eisen and Brown (1999; supra).

Sample total RNA (10 μg) samples are labeled using fluorescent Cy3 and Cy5 dyes. Labeled samples are resuspended in 4X SSC/0.03% SDS/4 μg salmon sperm DNA/2 μg tRNA/ 50mM Na-pyrophosphate, heated for 95°C for 2.5 minutes, spun down and placed on the array. The array is then covered with a glass coverslip and placed in a sealed chamber. The chamber is then kept in a water bath at 62°C overnight. The arrays are washed as described in Eisen and Brown (1999, *supra*) and scanned on a General Scanning 3000 laser scanner. The resulting files are subsequently quantified using IMAGENE, software (BioDiscovery, Los Angeles CA).

RT-PCR experiments may be performed to identify those genes induced after exposure to osmotic stress. Generally, the gene expression patterns from ground plant leaf tissue is examined. Reverse transcriptase PCR was conducted using gene specific primers within the coding region for each sequence identified. The primers were designed near the 3' region of each DNA binding sequence initially identified.

Total RNA from these ground leaf tissues was isolated using the CTAB extraction protocol. Once extracted total RNA was normalized in concentration across all the tissue types to ensure that the PCR reaction for each tissue received the same amount of cDNA template using the 28S band as reference. Poly(A+) RNA was purified using a modified protocol from the Qiagen OLIGOTEX purification kit batch protocol. cDNA was synthesized using standard protocols. After the first strand cDNA synthesis, primers for Actin 2 were used to normalize the concentration of cDNA across the tissue types. Actin 2 is found to be constitutively expressed in fairly equal levels across the tissue types we are investigating.

For RT PCR, cDNA template was mixed with corresponding primers and Taq DNA polymerase. Each reaction consisted of 0.2 μl cDNA template, 2 μl 10X Tricine buffer, 2 μl 10X Tricine buffer and 16.8 μl water, 0.05 μl Primer 1, 0.05 μl, Primer 2, 0.3 μl Taq DNA polymerase

and 8.6 µl water.

The 96 well plate is covered with microfilm and set in the thermocycler to start the reaction cycle. By way of illustration, the reaction cycle may comprise the following steps:

Step 1: 93° C for 3 min;

Step 2: 93° C for 30 sec;

Step 3: 65° C for 1 min;

Step 4: 72° C for 2 min;

Steps 2, 3 and 4 are repeated for 28 cycles;

Step 5: 72° C for 5 min; and

STEP 6 4° C.

To amplify more products, for example, to identify genes that have very low expression, additional steps may be performed: The following method illustrates a method that may be used in this regard. The PCR plate is placed back in the thermocycler for 8 more cycles of steps 2-4.

Step 2 93° C for 30 sec;

Step 3 65° C for 1 min;

Step 4 72° C for 2 min, repeated for 8 cycles; and

Step 5 4° C.

Eight microliters of PCR product and 1.5 µl of loading dye are loaded on a 1.2% agarose gel for analysis after 28 cycles and 36 cycles. Expression levels of specific transcripts are considered low if they were only detectable after 36 cycles of PCR. Expression levels are considered medium or high depending on the levels of transcript compared with observed transcript levels for an internal control such as actin2. Transcript levels are determined in repeat experiments and compared to transcript levels in control (e.g., non-transformed) plants.

Experiments were performed to identify those transformants or knockouts that exhibited an improved environmental stress tolerance. For such studies, the transformants were exposed to a variety of environmental stresses.

Germination assays all followed modifications of the same basic protocol. Sterile seeds were sown on the following conditional media. Plates were incubated at 22° C under 24-hour light (120-130 $\mu Ein/m^2/s$) in a growth chamber. Evaluation of germination and seedling vigor was conducted 3 to 15 days after planting. The basal media was 80% Murashige-Skoog medium (MS) + vitamins.

For salt and osmotic stress experiments, the medium was supplemented with 150 mM NaCl or 300 mM mannitol.

Carbon/nitrogen sensing experiments were conducted in basal media minus nitrogen plus 3% sucrose (-N) or in - basal media minus nitrogen plus 3% sucrose and 1 mM glutamine (N/+Gln).

Growth regulator sensitivity assays were performed in MS media, vitamins, and either 0.3 μM ABA, 9.4% sucrose 9.4%, or 5% glucose.

Temperature stress cold germination experiments were carried out at 8 °C. Heat stress germination experiments were conducted at 32 °C to 37° C for 6 hours of exposure.

For stress experiments conducted with more mature plants, seeds were germinated and grown for seven days on MS + vitamins + 1% sucrose at 22 °C and then transferred to chilling and heat stress conditions. The plants were either exposed to chilling stress (6 hour exposure to 4-8° C), or heat stress (32° C was applied for five days, after which the plants were transferred back 22 °C for recovery and evaluated after 5 days relative to controls not exposed to the depressed or elevated temperature).

high salt stress (6 hour exposure to 200 mM NaCl), drought stress (168 hours after removing water from trays), osmotic stress (6 hour exposure to 3 M mannitol), or nutrient limitation (nitrogen, phosphate, and potassium) (nitrogen: all components of MS medium remained constant except N was reduced to 20 mg/l of NH_4NO_3; phosphate: all components of MS medium except KH2PO4, which was replaced by K_2SO_4; potassium: all components of MS medium except removal of KNO_3 and KH_2PO_4, which were replaced by NaH_4PO_4).

Modified phenotypes observed for particular overexpressor or knockout plants are provided. For a particular overexpressor that shows a less beneficial characteristic, it may be more useful to select a plant with a decreased expression of the particular transcription factor. For a particular knockout that shows a less beneficial characteristic, it may be more useful to select a plant with an increased expression of the particular transcription factor.

The sequences of the Sequence Listing, can be used to prepare transgenic plants and plants with altered osmotic stress tolerance. The specific transgenic plants listed below are produced from the sequences of the Sequence Listing, as noted.

Example VIII: Genes that Confer Significant Improvements to Plants

Examples of genes and homologs that confer significant improvements to knockout or overexpressing plants are noted below. Experimental observations made by us with regard to specific genes whose expression has been modified in overexpressing or knock-out plants, and potential applications based on these observations, are also presented.

This example provides experimental evidence for increased biomass and abiotic stress tolerance controlled by the transcription factor polypeptides and polypeptides of the invention.

Salt stress assays are intended to find genes that confer better germination, seedling vigor or growth in high salt. Evaporation from the soil surface causes upward water movement and salt accumulation in the upper soil layer where the seeds are placed. Thus, germination normally takes

place at a salt concentration much higher than the mean salt concentration of in the whole soil profile. Plants differ in their tolerance to NaCl depending on their stage of development, therefore seed germination, seedling vigor, and plant growth responses are evaluated.

Osmotic stress assays (including NaCl and mannitol assays) are intended to determine if an osmotic stress phenotype is NaCl-specific or if it is a general osmotic stress related phenotype. Plants tolerant to osmotic stress could also have more tolerance to drought and/or freezing.

Drought assays are intended to find genes that mediate better plant survival after short-term, severe water deprivation. Ion leakage will be measured if needed. Osmotic stress tolerance would also support a drought tolerant phenotype.

Temperature stress assays are intended to find genes that confer better germination, seedling vigor or plant growth under temperature stress (cold, freezing and heat).

Sugar sensing assays are intended to find genes involved in sugar sensing by germinating seeds on high concentrations of sucrose and glucose and looking for degrees of hypocotyl elongation. The germination assay on mannitol controls for responses related to osmotic stress. Sugars are key regulatory molecules that affect diverse processes in higher plants including germination, growth, flowering, senescence, sugar metabolism and photosynthesis. Sucrose is the major transport form of photosynthate and its flux through cells has been shown to affect gene expression and alter storage compound accumulation in seeds (source-sink relationships). Glucose-specific hexose-sensing has also been described in plants and is implicated in cell division and repression of "famine" genes (photosynthetic or glyoxylate cycles).

Germination assays followed modifications of the same basic protocol. Sterile seeds were sown on the conditional media listed below. Plates were incubated at 22° C under 24-hour light (120-130 $\mu Ein/m^2/s$) in a growth chamber. Evaluation of germination and seedling vigor was conducted 3 to 15 days after planting. The basal media was 80% Murashige-Skoog medium (MS) + vitamins.

For salt and osmotic stress germination experiments, the medium was supplemented with 150 mM NaCl or 300 mM mannitol. Growth regulator sensitivity assays were performed in MS media, vitamins, and either 0.3 μM ABA, 9.4% sucrose, or 5% glucose.

Temperature stress cold germination experiments were carried out at 8 °C. Heat stress germination experiments were conducted at 32 °C to 37° C for 6 hours of exposure.

For stress experiments conducted with more mature plants, seeds were germinated and grown for seven days on MS + vitamins + 1% sucrose at 22 °C and then transferred to chilling and heat stress conditions. The plants were either exposed to chilling stress (6 hour exposure to 4-8° C), or heat stress (32° C was applied for five days, after which the plants were transferred back 22 °C for recovery and evaluated after 5 days relative to controls not exposed to the depressed or elevated temperature).

Results:

The overexpression of *A. thaliana* genes G481, G482, G485 and rice ortholog G3395 has been shown to increase osmotic stress tolerance. As noted below, changes in the activity of the G481 clade also produce alterations in flowering time.

G481 (Polynucleotide SEQ ID NO: 1)

Published information

G481 is equivalent to *AtHAP3a* which was identified by Edwards et al., ((1998) *Plant Physiol.* 117: 1015-1022) as an EST with extensive sequence homology to the yeast *HAP3*. Northern blot data from five different tissue samples indicates that G481 is primarily expressed in flower and/or silique, and root tissue. No other functional data is available for G481 in *Arabidopsis*.

Closely Related Genes from Other Species

There are several genes in the database from higher plants that show significant homology to G481 including, X59714 from corn, and two ESTs from tomato, AI486503 and AI782351.

Experimental Observations

The function of G481 was analyzed through its ectopic overexpression in plants. Except for darker color in one line (noted below), plants overexpressing G481 had a wild-type morphology. G481 overexpressors were found to be more tolerant to high sucrose and high salt (the latter is seen in Figure 8A), having better germination, longer radicles, and more cotyledon expansion. There was a consistent difference in the hypocotyl and root elongation in the overexpressor compared to wild-type controls. These results indicated that G481 is involved in sucrose-specific sugar sensing. Sucrose-sensing has been implicated in the regulation of source-sink relationships in plants.

In the T2 generation, one overexpressing line was darker green than wild-type plants, which may indicate a higher photosynthetic rate that would be consistent with the role of G481 in sugar sensing.

35S::G481 plants were also significantly larger and greener in a soil-based drought assay than wild-type controls plants After eight days of drought treatment overexpressing lines had a darker green and less withered appearance (Figure 7C) than those in the control group (Figure 7A). The differences in appearance between the control and G481-overexpressing plants after they were rewatered was even more striking. Eleven of twelve plants of this set of control plants died after rewatering (Figure 7B), indicating the inability to recover following severe water deprivation, whereas all nine of the

overexpressor plants of the line shown recovered from this drought treatment (Figure 7D). The results shown in Figures 7A-7D were typical of a number of control and 35S::G481-overexpressing lines.

One line of plants in which G481 was overexpressed under the control of the ARSK1 root-specific promoter was found to germinate better under cold conditions than wild-type plants.

Interestingly, in one *Arabidopsis* line in which G481 was knocked out, the plants were found to be more sensitive to high salt in a plate-based assay than wild-type plants, which indicates the importance of the role played by G481 in regulating osmotic stress tolerance, and demonstrates that the gene is both necessary and sufficient to fulfill that function.

A number of the 35S::G481 plants evaluated had a late flowering phenotype.

Utilities

The potential utility of G481 includes altering photosynthetic rate, which could also impact yield in vegetative tissues as well as seed. Sugars are key regulatory molecules that affect diverse processes in higher plants including germination, growth, flowering, senescence, sugar metabolism and photosynthesis. Sucrose is the major transport form of photosynthate and its flux through cells has been shown to affect gene expression and alter storage compound accumulation in seeds (source-sink relationships).

Since G481 overexpressing plants performed better than controls in drought experiments, this gene or its equivalogs may be used to improve seedling vigor, plant survival, as well as yield, quality, and range.

G482 (Polynucleotide SEQ ID NO: 3)

Published information

G482 is equivalent to *AtHAP3b* which was identified by Edwards et al. (1998) *Plant Physiol.* 117: 1015-1022) as an EST with homology to the yeast gene *HAP3b*. Their northern blot data suggests that *AtHAP3b* is expressed primarily in roots. No other functional information regarding G482 is publicly available.

Closely Related Genes from Other Species

The closest homology in the non-*Arabidopsis* plant database is within the B domain of G482, and therefore no potentially orthologous genes are available in the public domain.

Experimental Observations

RT-PCR analysis of endogenous levels of G482 transcripts indicated that this gene is expressed constitutively in all tissues tested. A cDNA array experiment supports the RT-PCR derived

tissue distribution data. G482 is not induced above basal levels in response to any environmental stress treatments tested.

A T-DNA insertion mutant for G482 was analyzed and was found to flower slightly later than control plants.

The function of G482 was also analyzed through its ectopic overexpression in plants. Plants overexpressing G482 had a wild-type morphology. Germination assays to measure salt tolerance demonstrated increased seedling growth when germinated on the high salt medium (Figure 8B).

35S::G482 transgenic plants also displayed an osmotic stress response phenotype similar to 35S::G481 transgenic lines. Five of ten overexpressing lines had increased seedling growth on medium containing 80% MS plus vitamins with 300 mM mannitol.

Three of ten 35S::G482 lines also demonstrated enhanced germination relative to controls after 6 h exposure to 32° C.

The majority of these 35S::G482 lines also demonstrated a slightly early flowering phenotype.

Utilities

The potential utilities of this gene include the ability to confer osmotic stress tolerance, as measured by salt, heat tolerance and improved germination in mannitol-containing media, during the germination stage of a crop plant. This would most likely impact survivability and yield. Evaporation of water from the soil surface causes upward water movement and salt accumulation in the upper soil layer, where the seeds are placed. Thus, germination normally takes place at a salt concentration much higher than the mean salt concentration in the whole soil profile.

Improved osmotic stress tolerance is also likely to result in enhanced seedling vigor, plant survival, improved yield, quality, and range. Osmotic stress assays, including subjecting plants to aqueous dissolved sugars, are often used as surrogate assays for improved water-stress (e.g., drought) response. Thus, G482 may also be used to improve plant performance under conditions of water deprivation, including increased seedling vigor, plant survival, yield, quality, and range.

G485 (Polynucleotide SEQ ID NO: 5)

Published Information

G485 is a member of the HAP3-like subfamily of CCAAT-box binding transcription factors. G485 corresponds to gene At4g14540, annotated by the *Arabidopsis* Genome Initiative. The gene corresponds to sequence 1042 from patent application WO0216655 (Harper et al. (2002)) on stress-regulated genes, transgenic plants and methods of use. In this application, G485 was reported to be cold responsive in their microarray analysis. No information is available about the function(s) of G485.

Experimental Observations

RT-PCR analyses of the endogenous levels of G485 indicated that this gene is expressed in all tissues and under all conditions tested.

A T-DNA insertion mutant for G485 was analyzed and was found to flower several days later than control plants (Figure 11A).

The effects of G485 overexpression were also studied. Interestingly, the gain of function and loss of function studies on G485 reveal opposing effects on flowering time. Under conditions of continuous light, approximately half of the 35S::G485 primary transformants flowered distinctly earlier than wild-type controls (up to a week sooner in 24-hour light) (Figure 11C). These effects were observed in each of two independent T1 plantings derived from separate transformation dates. Additionally, accelerated flowering was also seen in plants that overexpressed G485 from a two component system (35S::LexA;op-LexA::G485). These studies indicated that G485 is both sufficient to act as a floral activator, and is also necessary in that role within the plant. It should be noted that overexpression of G1820 (SEQ ID NO: 68), a member of the HAP5-like subfamily of CCAAT-box binding transcription factors had a similar effect on flowering time as G485. It is possible that G1820 interacts with G485 as part of a complex that binds and regulates the promoters of target genes involved in the regulation of flowering.

G485 overexpressor plants also matured and set siliques much more rapidly than wild type controls (Figure 11B).

G485 overexpressing plants were shown to have enhanced response to stress-related treatments in plate-based germination assays. As seen in Figures 10A-10D and Table 6, 35S::G485 lines showed enhanced cotyledon expansion and root growth seen in the overexpressing seedlings in cold, high sucrose, high salt and ABA treatments, as compared to wild-type controls with the same treatments seen in Figures 10E-10H.

Utilities

Based on the loss of function and gain of function phenotypes, G485 could be used to modify flowering time.

The delayed flowering displayed by G485 knockouts suggests that the gene might be used to manipulate the flowering time of commercial species. In particular, an extension of vegetative growth can significantly increase biomass and result in substantial yield increases. In some species (for example sugar beet), where the vegetative parts of the plant constitute the crop, it would be advantageous to delay or suppress flowering in order to prevent resources being diverted into reproductive development. Additionally, delaying flowering beyond the normal time of harvest could

alleviate the risk of transgenic pollen escape from such crops.

The early flowering effects see in the G485 overexpressors could be applied to accelerate flowering, or eliminate any requirement for vernalization. In some instances, a faster cycling time might allow additional harvests of a crop to be made within a given growing season. Shortening generation times could also help speed-up breeding programs, particularly in species such as trees, which typically grow for many years before flowering.

Table 6 provides a summary of the data collected from one series of experiments conducted with plants overexpressing G482 or a paralog of G482. In each case the promoter used for regulating the introduced transcription factor was the cauliflower mosaic virus 35S transcription initiation region. The column headings include the transcription factors used to transform the *Arabidopsis* plants listed by Gene ID (GID) numbers, the corresponding polypeptide SEQ ID NO; the project type indicating the nature of the promoter-gene interaction, and the ratio of lines determine to have one of the enhanced abiotic stress phenotypes listed over the number of lines tested

G3395 (Polynucleotide SEQ ID NO: 73)

Published Information

G3395, an ortholog of G482, is a member of the HAP3-like subfamily of CCAAT-box binding transcription factors. G3395 corresponds to polypeptide BAC76331 ("NF-YB subunit of rice").

Closely Related Genes from Other Species

The most closely related gene sequence found in GenBank appears to be the nearly identical AB095438 ("OsNF-YB2 mRNA for NF-YB").

Experimental Observations

The function of G3395 was analyzed through its ectopic overexpression in plants. One of the lines of G3395 overexpressors tested was found to be more tolerant to high salt levels, producing larger and greener seedlings in a high salt germination assay.

Utilities

The potential utilities of this gene include the ability to confer osmotic stress tolerance, particularly during the germination stage of a crop plant.

Table 6. Summary of Results of Physiological Assays.

GID	Polypeptide SEQ ID NO	Promoter	One or two Component Transformation Type	Overexpressor lines showing phenotype					
				Heat tolerance	Drought tolerance	Improved germ. in high NaCl	Improved germ. in high sugar	ABA sens.	Improved germ. in cold
G482	2	CaMV 35S	2-components-supTfn	+			+*		
		CaMV 35S	Direct promoter-fusion			+			
G481	4	CaMV 35S	Direct promoter-fusion		+		++		
		ARSK1	2-components-supTfn						++
		CaMV 35S	Superactivation			+			
		CaMV 35S	RNAi (GS)	++	+		+**		
G485	6	CaMV 35S	2-components-supTfn			+	+**	+	+
G3395	74	CaMV 35S	Direct promoter-fusion			+			

* Mannitol ** Sucrose

Abbreviations: Sens. Sensitivity Germ. Germination

\+ Moderate trait manifestation in one or more lines tested

++ Strong trait manifestation in one or more lines tested

EXAMPLE IX. CCAAT family transcription factors and flowering time

We have also found that overexpressed CCAAT genes also have a highly noticeable effect on the timing of onset of flowering. G482 (SEQ ID NO: 3), G485 (SEQ ID NO: 5), G1248 (SEQ ID NO: 69), G1781 (SEQ ID NO: 71) and related crop orthologs G3398 (SEQ ID NO: 75), G3435 (SEQ ID NO: 47), and G3436 (SEQ ID NO: 49), accelerate onset of flowering when overexpressed in *Arabidopsis*.

Conversely, overexpression of G481, G1364 and related crop orthologs G3471 (SEQ ID NO: 23), G3434 (SEQ ID NO: 77), and G3395 (SEQ ID NO: 73), produce a slight but reproducible delay in flowering in *Arabidopsis*. Results of knockout and RNAi studies confirm these findings. Knocked-out G485 and G482 plants exhibit a delay in flowering, and RNAi lines (using a construct designed to knock-out any member of the clade) are late flowering.

Thus, it appears that genes in the node of the tree clustered around G481 act to repress flowering, whereas those clustered around G482 and G485 act to promote flowering.

Interestingly, the addition of an activation domain appears to convert a floral repressor to a floral activator. Overexpression of a fusion protein comprising G481 fused at its carboxyl end with a GAL4 activation domain causes early flowering that is comparable to the effects caused by G482 or G482 overexpression.

An alignment of some of these HAP3 genes, seen in Figures 6A-6F, shows the high degree of

conservation within the B domain, particularly in the B domain extending from Figure 6B through Figure 6C. These proteins are almost identical within the B domain, but the composition of two residue positions within the B domain correlates with effects of expression on flowering. These positions are indicated by arrows in Figure 6B. The residue position indicated by the downward-pointing arrow in Figure 6B is a serine residue in G1364, G2345 and G481 and a glycine residue in G482 and G485. The composition at this position correlate with flowering time when the polypeptide is overexpressed. The former group with a serine residue at this position induces late flowering when overexpressed, whereas the latter group with the glycine residue is distinguished by very early flowering upon overexpression. This study was expanded to include other polypeptides of the HAP3 family that compared the effects on flowering time and the relationship to the serine/glycine residue, including orthologous soy, corn and rice polypeptides. In each case, a glycine present at this position was associated with early flowering, and a serine residue was associated with a delay in flowering (G486 was found to possess a cysteine residue at this position, and one overexpressing T2 line appeared to have a late flowering phenotype). Similar observations were made with respect the other residue position, as indicated by the upward-pointing arrow in Figure 6B) where orthologous polypeptides that cause late flowering, including soy, corn and rice polypeptides, possess a glycine or alanine residue at this position, and orthologs derived these species that produce an early flowering phenotype have a serine residue at the position. These results suggest that these residue positions are essential for determining whether these polypeptides are able to interact effectively with their partners in the multi-subunit complex and bind effectively to a promoter CCAAT box.

Example X: Identification of Homologous Sequences

This example describes identification of genes that are orthologous to *Arabidopsis thaliana* transcription factors from a computer homology search.

Homologous sequences, including those of paralogs and orthologs from *Arabidopsis* and other plant species, were identified using database sequence search tools, such as the Basic Local Alignment Search Tool (BLAST) (Altschul et al. (1990) *J. Mol. Biol.* 215: 403-410; and Altschul et al. (1997) *Nucleic Acid Res.* 25: 3389-3402). The tblastx sequence analysis programs were employed using the BLOSUM-62 scoring matrix (Henikoff and Henikoff (1992) *Proc .Natl. Acad. Sci.* 89: 10915-10919). The entire NCBI GenBank database was filtered for sequences from all plants except *Arabidopsis thaliana* by selecting all entries in the NCBI GenBank database associated with NCBI taxonomic ID 33090 (Viridiplantae; all plants) and excluding entries associated with taxonomic ID 3701 (*Arabidopsis thaliana*).

These sequences are compared to sequences SEQ ID NOs: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21,

23, 25, 27, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93 or polynucleotides that encode polypeptide SEQ ID NOs: 29-32, using the Washington University TBLASTX algorithm (version 2.0a19MP) at the default settings using gapped alignments with the filter "off". For each these genes, individual comparisons were ordered by probability score (P-value), where the score reflects the probability that a particular alignment occurred by chance. For example, a score of 3.6e-40 is 3.6 x 10-40. In addition to P-values, comparisons were also scored by percentage identity. Percentage identity reflects the degree to which two segments of DNA or protein are identical over a particular length. Examples of sequences so identified are presented in Table 5. The percent sequence identity among these sequences can be as low as 49%, or even lower sequence identity.

Candidate paralogous sequences were identified among *Arabidopsis* transcription factors through alignment, identity, and phylogenic relationships. Paralogs of G481 so determined include G482, G485, G1364, and G2345. Candidate orthologous sequences were identified from proprietary unigene sets of plant gene sequences in *Zea mays*, *Glycine max* and *Oryza sativa* based on significant homology to *Arabidopsis* transcription factors. These candidates were reciprocally compared to the set of *Arabidopsis* transcription factors. If the candidate showed maximal similarity in the protein domain to the eliciting transcription factor or to a paralog of the eliciting transcription factor, then it was considered to be an ortholog. Identified non-*Arabidopsis* sequences that were shown in this manner to be orthologous to the *Arabidopsis* sequences are provided in Table 5.

Example XI: Screen of Plant cDNA library for Sequence Encoding a Transcription Factor DNA Binding Domain That Binds To a Transcription Factor Binding Promoter Element and Demonstration of Protein Transcription Regulation Activity.

The "one-hybrid" strategy (Li and Herskowitz (1993) *Science* 262: 1870-1874) is used to screen for plant cDNA clones encoding a polypeptide comprising a transcription factor DNA binding domain, a conserved domain. In brief, yeast strains are constructed that contain a lacZ reporter gene with either wild-type or mutant transcription factor binding promoter element sequences in place of the normal UAS (upstream activator sequence) of the GALL promoter. Yeast reporter strains are constructed that carry transcription factor binding promoter element sequences as UAS elements are operably linked upstream (5') of a lacZ reporter gene with a minimal GAL1 promoter. The strains are transformed with a plant expression library that contains random cDNA inserts fused to the GAL4 activation domain (GAL4-ACT) and screened for blue colony formation on X-gal-treated filters (X-gal: 5-bromo-4-chloro-3-indolyl-ß-D-galactoside; Invitrogen Corporation, Carlsbad CA). Alternatively, the strains are transformed with a cDNA polynucleotide encoding a known transcription factor DNA

binding domain polypeptide sequence.

Yeast strains carrying these reporter constructs produce low levels of beta-galactosidase and form white colonies on filters containing X-gal. The reporter strains carrying wild-type transcription factor binding promoter element sequences are transformed with a polynucleotide that encodes a polypeptide comprising a plant transcription factor DNA binding domain operably linked to the acidic activator domain of the yeast GAL4 transcription factor, "GAL4-ACT". The clones that contain a polynucleotide encoding a transcription factor DNA binding domain operably linked to GLA4-ACT can bind upstream of the lacZ reporter genes carrying the wild-type transcription factor binding promoter element sequence, activate transcription of the lacZ gene and result in yeast forming blue colonies on X-gal-treated filters.

Upon screening about 2 x 10^6 yeast transformants, positive cDNA clones are isolated; i.e., clones that cause yeast strains carrying lacZ reporters operably linked to wild-type transcription factor binding promoter elements to form blue colonies on X-gal-treated filters. The cDNA clones do not cause a yeast strain carrying a mutant type transcription factor binding promoter elements fused to LacZ to turn blue. Thus, a polynucleotide encoding transcription factor DNA binding domain, a conserved domain, is shown to activate transcription of a gene.

Example XII: Gel Shift Assays.

The presence of a transcription factor comprising a DNA binding domain which binds to a DNA transcription factor binding element is evaluated using the following gel shift assay. The transcription factor is recombinantly expressed and isolated from *E. coli* or isolated from plant material. Total soluble protein, including transcription factor, (40 ng) is incubated at room temperature in 10 µl of 1 x binding buffer (15 mM HEPES (pH 7.9), 1 mM EDTA, 30 mM KCl, 5% glycerol, 5% bovine serum albumin, 1 mM DTT) plus 50 ng poly(dl-dC):poly(dl-dC) (Pharmacia, Piscataway NJ) with or without 100 ng competitor DNA. After 10 minutes incubation, probe DNA comprising a DNA transcription factor binding element (1 ng) that has been ^{32}P-labeled by end-filling (Sambrook et al. (1989) *supra*) is added and the mixture incubated for an additional 10 minutes. Samples are loaded onto polyacrylamide gels (4% w/v) and fractionated by electrophoresis at 150V for 2h (Sambrook et al. *supra*). The degree of transcription factor-probe DNA binding is visualized using autoradiography. Probes and competitor DNAs are prepared from oligonucleotide inserts ligated into the BamHI site of pUC118 (Vieira et al. (1987) *Methods Enzymol.* 153: 3-11). Orientation and concatenation number of the inserts are determined by dideoxy DNA sequence analysis (Sambrook et al. *supra*). Inserts are recovered after restriction digestion with EcoRI and HindIII and fractionation on polyacrylamide gels (12% w/v) (Sambrook et al. *supra*).

Example XIII. Introduction of Polynucleotides into Dicotyledonous Plants

SEQ ID NOs: 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, polynucleotides that encode polypeptide SEQ ID NOs: 29-32, paralogous, and orthologous sequences recombined into pMEN20 or pMEN65 expression vectors are transformed into a plant for the purpose of modifying plant traits. The cloning vector may be introduced into a variety of cereal plants by means well known in the art such as, for example, direct DNA transfer or *Agrobacterium tumefaciens*-mediated transformation. It is now routine to produce transgenic plants using most dicot plants (see Weissbach and Weissbach, (1989*) supra;* Gelvin et al. (1990) *supra*; Herrera-Estrella et al. (1983) *supra*; Bevan (1984) *supra*; and Klee (1985) *supra*). Methods for analysis of traits are routine in the art and examples are disclosed above.

Example XIV: Transformation of Cereal Plants with an Expression Vector

Cereal plants such as, but not limited to, corn, wheat, rice, sorghum, or barley, may also be transformed with the present polynucleotide sequences in pMEN20 or pMEN65 expression vectors for the purpose of modifying plant traits. For example, pMEN020 may be modified to replace the NptII coding region with the BAR gene of *Streptomyces hygroscopicus* that confers resistance to phosphinothricin. The KpnI and BglII sites of the Bar gene are removed by site-directed mutagenesis with silent codon changes.

The cloning vector may be introduced into a variety of cereal plants by means well known in the art such as, for example, direct DNA transfer or *Agrobacterium tumefaciens*-mediated transformation. It is now routine to produce transgenic plants of most cereal crops (Vasil (1994) *Plant Mol. Biol.* 25: 925-937) such as corn, wheat, rice, sorghum (Cassas et al. (1993) *Proc. Natl. Acad. Sci.* 90: 11212-11216, and barley (Wan and Lemeaux (1994) *Plant Physiol.* 104:37-48. DNA transfer methods such as the microprojectile can be used for corn (Fromm et al. (1990) *Bio/Technol.* 8: 833-839); Gordon-Kamm et al. (1990) *Plant Cell* 2: 603-618; Ishida (1990) *Nature Biotechnol.* 14:745-750), wheat (Vasil et al. (1992) *Bio/Technol.* 10:667-674; Vasil et al. (1993) *Bio/Technol.* 11:1553-1558; Weeks et al. (1993) *Plant Physiol.* 102:1077-1084), rice (Christou (1991) *Bio/Technol.* 9:957-962; Hiei et al. (1994) *Plant J.* 6:271-282; Aldemita and Hodges (1996) *Planta* 199:612-617; and Hiei et al. (1997) *Plant Mol. Biol.* 35:205-218). For most cereal plants, embryogenic cells derived from immature scutellum tissues are the preferred cellular targets for transformation (Hiei et al. (1997) *Plant Mol. Biol.* 35:205-218; Vasil (1994) *Plant Mol. Biol.* 25: 925-937).

Vectors according to the present invention may be transformed into corn embryogenic cells derived from immature scutellar tissue by using microprojectile bombardment, with the A188XB73

genotype as the preferred genotype (Fromm et al. (1990) *Bio/Technol.* 8: 833-839; Gordon-Kamm et al. (1990) *Plant Cell* 2: 603-618). After microprojectile bombardment the tissues are selected on phosphinothricin to identify the transgenic embryogenic cells (Gordon-Kamm et al. (1990) *Plant Cell* 2: 603-618). Transgenic plants are regenerated by standard corn regeneration techniques (Fromm et al. (1990) *Bio/Technol.* 8: 833-839; Gordon-Kamm et al. (1990) *Plant Cell* 2: 603-618).

The plasmids prepared as described above can also be used to produce transgenic wheat and rice plants (Christou (1991) *Bio/Technol.* 9:957-962; Hiei et al. (1994) *Plant J.* 6:271-282; Aldemita and Hodges (1996) *Planta* 199:612-617; and Hiei et al. (1997) *Plant Mol. Biol.* 35:205-218) that coordinately express genes of interest by following standard transformation protocols known to those skilled in the art for rice and wheat (Vasil et al. (1992) *Bio/Technol.* 10:667-674; Vasil et al. (1993) *Bio/Technol.* 11:1553-1558; and Weeks et al. (1993) *Plant Physiol.* 102:1077-1084), where the bar gene is used as the selectable marker.

Example XV: Genes that Confer Significant Improvements to non-*Arabidopsis* species

The function of orthologs of G481 and G482 may be analyzed through their ectopic overexpression in plants using the CaMV 35S or other appropriate promoter, as identified above. These genes encode members of the HAP3 subfamily of CCAAT-box binding transcription factors and include those found in Table 5, Figures 3 and 4, and, for example, polynucleotide sequences from *Arabidopsis thaliana* (SEQ ID NO: 1, 3, 5, 7, 9, 69, and 71), *Glycine max* (SEQ ID NO: 11, 13, 15, 17, 19, 21, 23, 25, 51, 79, 81, 83, and 85), *Solanum tuberosum* (BQ505706), *Medicago truncatula* (AC122165), *Lycopersicon esculentum* (SEQ ID NO: 63, SEQ ID NO: 65, and BG642751), *Rosa* hybrid (BQ104671), *Poncirus trifoliata* (CD574709), *Populus balsamifera* subsp. *trichocarpa* (BU880488), *Zea mays* (SEQ ID NO: 33, 35, 37, 39, 41, 43, 45, 47, 49, 55, 57, 77, 93, CC429501; and AX584265), *Oryza sativa* (SEQ ID NO: 27, 53, 73, 75, 87, 89, AAAA01003638, AP005193, AC108500, AP004366, AP003266, AP004179, AC104284, and AP120529), and *Triticum aestivum* (SEQ ID NO: 59, 61, and BJ248969). The function of specific HAP3 subfamily of CCAAT-box binding transcription factor genes that may be analyzed through ectopic overexpression in plants also includes rice nucleic acid sequences that encode polypeptides SEQ ID NO: 29-32, corn sequence gi115840, and wheat sequence gi16902058. These polynucleotide and polypeptide sequences derived from monocots may be used to transform both monocot and dicot plants, and those derived from dicots may also be used to transform either group, although some of these sequences will function best if the gene is transformed into the a plant from the same group as that from which the sequence is derived.

Seeds of these transgenic plants are subjected to germination assays to measure sucrose sensing. Sterile monocot seeds, including, but not limited to, corn, rice, wheat, rye and sorghum, as

well as dicots including, but not limited to soybean and alfalfa, are sown on 80% MS medium plus vitamins with 9.4% sucrose; control media lack sucrose. All assay plates are then incubated at 22° C under 24-hour light, 120-130 μEin/m^2/s, in a growth chamber. Evaluation of germination and seedling vigor is then conducted three days after planting. Overexpressors of these genes may be found to be more tolerant to high sucrose by having better germination, longer radicles, and more cotyledon expansion. These results would indicate that overexpressors of G482 orthologs are involved in sucrose-specific sugar sensing.

Plants overexpressing G482 orthologs may also be subjected to soil-based drought assays to identify those lines that are more tolerant to water deprivation than wild-type control plants. Generally, 35S: G482 ortholog overexpressing plants will appear significantly larger and greener, with less wilting or desiccation, than wild-type controls plants, particularly after a period of water deprivation is followed by rewatering and a subsequent incubation period.

Example XVI: Identification of Orthologous and Paralogous Sequences

Orthologs to *Arabidopsis* genes may identified by several methods, including hybridization, amplification, or bioinformatically. This example describes how one may identify homologs to the *Arabidopsis* AP2 family transcription factor CBF1 (polynucleotide SEQ ID NO: 95, encoded polypeptide SEQ ID NO: 96), which confers tolerance to abiotic stresses (Thomashow et al. (2002) US Patent No. 6,417,428), and an example to confirm the function of homologous sequences. In this example, orthologs to CBF1 were found in canola (*Brassica napus*) using polymerase chain reaction (PCR).

Degenerate primers were designed for regions of AP2 binding domain and outside of the AP2 (carboxyl terminal domain):

Mol 368 (reverse) 5'- CAY CCN ATH TAY MGN GGN GT -3' (SEQ ID NO: 103)

Mol 378 (forward) 5'- GGN ARN ARC ATN CCY TCN GCC -3' (SEQ ID NO: 104)

(Y: C/T, N: A/C/G/T, H: A/C/T, M: A/C, R: A/G)

Primer Mol 368 is in the AP2 binding domain of CBF1 (amino acid sequence: His-Pro-Ile-Tyr-Arg-Gly-Val) while primer Mol 378 is outside the AP2 domain (carboxyl terminal domain) (amino acid sequence: Met-Ala-Glu-Gly-Met-Leu-Leu-Pro).

The genomic DNA isolated from *B. napus* was PCR-amplified by using these primers following these conditions: an initial denaturation step of 2 min at 93° C; 35 cycles of 93° C for 1 min, 55° C for 1 min, and 72° C for 1 min ; and a final incubation of 7 min at 72° C at the end of cycling.

The PCR products were separated by electrophoresis on a 1.2% agarose gel and transferred to nylon membrane and hybridized with the AT CBF1 probe prepared from *Arabidopsis* genomic DNA by PCR amplification. The hybridized products were visualized by colorimetric detection system (Boehringer Mannheim) and the corresponding bands from a similar agarose gel were isolated using the Qiagen Extraction Kit (Qiagen). The DNA fragments were ligated into the TA clone vector from TOPO TA Cloning Kit (Invitrogen) and transformed into *E. coli* strain TOP10 (Invitrogen).

Seven colonies were picked and the inserts were sequenced on an ABI 377 machine from both strands of sense and antisense after plasmid DNA isolation. The DNA sequence was edited by sequencer and aligned with the AtCBF1 by GCG software and NCBI blast searching.

The nucleic acid sequence and amino acid sequence of one canola ortholog found in this manner (bnCBF1; polynucleotide SEQ ID NO: 101 and polypeptide SEQ ID NO: 102) identified by this process is shown in the Sequence Listing.

The aligned amino acid sequences show that the bnCBF1 gene has 88% identity with the *Arabidopsis* sequence in the AP2 domain region and 85% identity with the *Arabidopsis* sequence outside the AP2 domain when aligned for two insertion sequences that are outside the AP2 domain.

Similarly, paralogous sequences to *Arabidopsis* genes, such as *CBF1*, may also be identified.

Two paralogs of CBF1 from *Arabidopsis thaliana*: *CBF2* and *CBF3*. *CBF2* and *CBF3* have been cloned and sequenced as described below. The sequences of the DNA SEQ ID NO: 97 and 99 and encoded proteins SEQ ID NO: 98 and 100 are set forth in the Sequence Listing.

A lambda cDNA library prepared from RNA isolated from *Arabidopsis thaliana* ecotype Columbia (Lin and Thomashow (1992) *Plant Physiol.* 99: 519-525) was screened for recombinant clones that carried inserts related to the *CBF1* gene (Stockinger et al. (1997) *Proc. Natl. Acad. Sci.* 94:1035-1040). CBF1 was ^{32}P-radiolabeled by random priming (Sambrook et al. *supra*) and used to screen the library by the plaque-lift technique using standard stringent hybridization and wash conditions (Hajela et al. (1990) *Plant Physiol.* 93:1246-1252; Sambrook et al. *supra*) 6 X SSPE buffer, 60° C for hybridization and 0.1 X SSPE buffer and 60° C for washes). Twelve positively hybridizing clones were obtained and the DNA sequences of the cDNA inserts were determined. The results indicated that the clones fell into three classes. One class carried inserts corresponding to *CBF1*. The two other classes carried sequences corresponding to two different homologs of *CBF1*, designated *CBF2* and *CBF3*. The nucleic acid sequences and predicted protein coding sequences for *Arabidopsis CBF1*, *CBF2* and *CBF3* are listed in the Sequence Listing (SEQ ID NOs:95, 97, 99 and SEQ ID NOs: 96, 98, and 100, respectively). The nucleic acid sequences and predicted protein coding sequence for *Brassica napus CBF* ortholog is listed in the Sequence Listing (SEQ ID NOs: 101 and 102, respectively).

A comparison of the nucleic acid sequences of *Arabidopsis CBF1*, *CBF2* and *CBF3* indicate that they are 83 to 85% identical as shown in Table 7.

TABLE 7

	Percent identity[a]	
	DNA[b]	Polypeptide
cbf1/cbf2	85	86
cbf1/cbf3	83	84
cbf2/cbf3	84	85

[a] Percent identity was determined using the *Clustal* algorithm from the Megalign program (DNASTAR, Inc.).

[b] Comparisons of the nucleic acid sequences of the open reading frames are shown.

Similarly, the amino acid sequences of the three CBF polypeptides range from 84 to 86% identity. An alignment of the three amino acidic sequences reveals that most of the differences in amino acid sequence occur in the acidic C-terminal half of the polypeptide. This region of CBF1 serves as an activation domain in both yeast and *Arabidopsis* (not shown).

Residues 47 to 106 of CBF1 correspond to the AP2 domain of the protein, a DNA binding motif that to date, has only been found in plant proteins. A comparison of the AP2 domains of CBF1, CBF2 and CBF3 indicates that there are a few differences in amino acid sequence. These differences in amino acid sequence might have an effect on DNA binding specificity.

Example XVII: Transformation of Canola with a Plasmid Containing CBF1, CBF2, or CBF3

After identifying homologous genes to CBF1, canola was transformed with a plasmid containing the *Arabidopsis* CBF1, CBF2, or CBF3 genes cloned into the vector pGA643 (An (1987) *Methods Enzymol.* 253: 292). In these constructs the CBF genes were expressed constitutively under the CaMV 35S promoter. In addition, the CBF1 gene was cloned under the control of the *Arabidopsis* COR15 promoter in the same vector pGA643. Each construct was transformed into *Agrobacterium* strain GV3101. Transformed Agrobacteria were grown for 2 days in minimal AB medium containing appropriate antibiotics.

Spring canola (*B. napus* cv. Westar) was transformed using the protocol of Moloney et al. ((1989) *Plant Cell Reports* 8: 238) with some modifications as described. Briefly, seeds were sterilized and plated on half strength MS medium, containing 1% sucrose. Plates were incubated at 24° C under 60-80 $\mu E/m^2s$ light using a16 hour light/ 8 hour dark photoperiod. Cotyledons from 4-5 day old seedlings were collected, the petioles cut and dipped into the *Agrobacterium* solution. The dipped cotyledons were placed on co-cultivation medium at a density of 20 cotyledons/plate and incubated as described above for 3 days. Explants were transferred to the same media, but containing 300 mg/l

timentin (SmithKline Beecham, PA) and thinned to 10 cotyledons/plate. After 7 days explants were transferred to Selection/Regeneration medium. Transfers were continued every 2-3 weeks (2 or 3 times) until shoots had developed. Shoots were transferred to Shoot-Elongation medium every 2-3 weeks. Healthy looking shoots were transferred to rooting medium. Once good roots had developed, the plants were placed into moist potting soil.

The transformed plants were analyzed for the presence of the NPTII gene/ kanamycin resistance by ELISA, using the ELISA NPTII kit from 5Prime-3Prime Inc. (Boulder, CO). Approximately 70% of the screened plants were NPTII positive; these plants were further analyzed.

From Northern blot analysis of the plants that were transformed with the constitutively expressing constructs, showed expression of the CBF genes and all CBF genes were capable of inducing the *Brassica napus* cold-regulated gene BN115 (homolog of the *Arabidopsis* COR15 gene). Most of the transgenic plants appear to exhibit a normal growth phenotype. As expected, the transgenic plants are more freezing tolerant than the wild-type plants. Using the electrolyte leakage of leaves test, the control showed a 50% leakage at –2 to –3° C. Spring canola transformed with either CBF1 or CBF2 showed a 50% leakage at –6 to –7° C. Spring canola transformed with CBF3 shows a 50% leakage at about –10 to –15° C. Winter canola transformed with CBF3 may show a 50% leakage at about –16 to –20° C. Furthermore, if the spring or winter canola are cold acclimated the transformed plants may exhibit a further increase in freezing tolerance of at least –2° C.

To test salinity tolerance of the transformed plants, plants were watered with 150 mM NaCl. Plants overexpressing CBF1, CBF2 or CBF3 grew better compared with plants that had not been transformed with CBF1, CBF2 or CBF3.

These results demonstrate that homologs of *Arabidopsis* transcription factors can be identified and shown to confer similar functions in non-*Arabidopsis* plant species.

Example XVIII: Cloning of transcription factor promoters

Promoters are isolated from transcription factor genes that have gene expression patterns useful for a range of applications, as determined by methods well known in the art (including transcript profile analysis with cDNA or oligonucleotide microarrays, Northern blot analysis, semi-quantitative or quantitative RT-PCR). Interesting gene expression profiles are revealed by determining transcript abundance for a selected transcription factor gene after exposure of plants to a range of different experimental conditions, and in a range of different tissue or organ types, or developmental stages. Experimental conditions to which plants are exposed for this purpose includes cold, heat, drought, osmotic challenge, varied hormone concentrations (ABA, GA, auxin, cytokinin, salicylic acid, brassinosteroid), pathogen and pest challenge. The tissue types and developmental stages include stem,

root, flower, rosette leaves, cauline leaves, siliques, germinating seed, and meristematic tissue. The set of expression levels provides a pattern that is determined by the regulatory elements of the gene promoter.

Transcription factor promoters for the genes disclosed herein are obtained by cloning 1.5 kb to 2.0 kb of genomic sequence immediately upstream of the translation start codon for the coding sequence of the encoded transcription factor protein. This region includes the 5'-UTR of the transcription factor gene, which can comprise regulatory elements. The 1.5 kb to 2.0 kb region is cloned through PCR methods, using primers that include one in the 3' direction located at the translation start codon (including appropriate adaptor sequence), and one in the 5' direction located from 1.5 kb to 2.0 kb upstream of the translation start codon (including appropriate adaptor sequence). The desired fragments are PCR-amplified from *Arabidopsis* Col-0 genomic DNA using high-fidelity Taq DNA polymerase to minimize the incorporation of point mutation(s). The cloning primers incorporate two rare restriction sites, such as Not1 and Sfi1, found at low frequency throughout the *Arabidopsis* genome. Additional restriction sites are used in the instances where a Not1 or Sfi1 restriction site is present within the promoter.

The 1.5-2.0 kb fragment upstream from the translation start codon, including the 5'-untranslated region of the transcription factor, is cloned in a binary transformation vector immediately upstream of a suitable reporter gene, or a transactivator gene that is capable of programming expression of a reporter gene in a second gene construct. Reporter genes used include green fluorescent protein (and related fluorescent protein color variants), beta-glucuronidase, and luciferase. Suitable transactivator genes include LexA-GAL4, along with a transactivatable reporter in a second binary plasmid (as disclosed in US patent application 09/958,131, incorporated herein by reference). The binary plasmid(s) is transferred into *Agrobacterium* and the structure of the plasmid confirmed by PCR. These strains are introduced into *Arabidopsis* plants as described in other examples, and gene expression patterns determined according to standard methods know to one skilled in the art for monitoring GFP fluorescence, beta-glucuronidase activity, or luminescence.

All publications and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication or patent application was specifically and individually indicated to be incorporated by reference.

The present invention is not limited by the specific embodiments described herein. The invention now being fully described, it will be apparent to one of ordinary skill in the art that many changes and modifications can be made thereto without departing from the spirit or scope of the appended claims. Modifications that become apparent from the foregoing description and accompanying figures fall within the scope of the claims.